# Process View for a Data Stream Processing Strategy based on Measurement Metadata

Mario Diván[1,2], Luis Olsina[2]

[1]Facultad de Ciencias Económicas y Jurídicas, UNLPam, Santa Rosa, La Pampa, Argentina
[2]Facultad de Ingeniería, UNLPam, General Pico, La Pampa, Argentina.

{mjdivan,olsinal}@ing.unlpam.edu.ar

**Abstract.** The data stream processing strategy based on measurement metadata, is an enhanced approach that relies on a measurement and evaluation conceptual framework. The strategy uses measures (data sets) and their linked metadata to incorporate detective and predictive behavior as well as to trigger alarms in a proactive way. This work discusses the process formalization for its data stream processing components, and their interactions. The process specification is based on the SPEM language, and the main specified activities ranges from data source configuration of heterogeneous data sources, to statistical analysis and classifiers for decision making. As result, the proposed strategy becomes more consistent, repeatable and communicable from the process modeling viewpoint. Also, to illustrate the approach, excerpts of a developed outpatient monitoring system are used.

**Keywords:** Data Stream, Process, Measure, Metadata, C-INCAMI, SPEM

## 1 Introduction

Nowadays, there are software systems which make customized processing of data sets, generated in a continuous way (i.e. streams) in response to queries and/or to adjust their behavior depending on the arrival of new data [1]. Practical examples of those applications are aimed for instance at monitoring vital parameters of patients; behavioral tracking of financial markets; monitoring of air traffic, amongst others. In this kind of applications, the arrival of a new data represents receiving a value (e.g. for a cardiac frequency, a foreign currency rate, etc.) associated to a syntactical behavior. Often, many current applications just analyze the number (value) itself without a semantic support, disregarding not only the metadata of measures, but also the context in which the phenomenon occurs. In order to understand the meaning of arriving data and then act accordingly, such applications might incorporate a logic layer, i.e. procedures and metadata, which transforms and/or interpret data streams. Due to a lack of

clear separation of concerns between the syntactic and semantic aspects of those applications, very often an expert (e.g., a doctor responsible for the monitoring of the outpatient system) should intervene in order to interpret the situation. Therefore, we argue that given the state-of-the-art of IT in metadata and semantic processing, the intervention of experts should be minimized as long as the applications can perform the job. In fact, our proposed Strategy for Data Stream Processing based on Measurement Metadata (SDSPbMM) [2] supports these concerns.

Regarding the semantic ground for measurement and evaluation (M&E), the C-INCAMI (*Context-Information Need*, *Concept model*, *Attribute*, *Metric* and *Indicator*) conceptual framework [3,4] is built on an ontology which includes the concepts and relationships needed to specify data and metadata for any M&E project. So, SDSPbMM incorporates metadata to the measurement process using the C-INCAMI framework in order to promote consistence and comparability of results. Unlike other data stream processing strategies [5,6,7], SDSPbMM is able to support the appropriate processing of measures generated from heterogeneous data sources thanks to the included metadata. In this way, each measure is analyzed considering its semantic and context as per the formal definition of each M&E project.

Therefore, the arrival of a measure -which can be a measured value of an attribute of a target entity, or a context property value of a target entity's situation- is not just a value because the associated metadata allows to guide the stream processing coherently from their respective meanings in the M&E project. So far, the SDSPbMM strategy was described basically regarding its components [8,9]. But its process specification had been left apart, so it was difficult to understand the specific task sequences, parallelisms, iterations, among other aspects. Thus, the specific contributions of this paper, by using the SPEM (*Software & Systems Process Engineering Metamodel*) [10] language are related to: *i) data source configuration*: the formalization of the configuration process which establishes a mapping among data sources associated with a measurement adapter and the metrics associated with attributes of a target entity and its context properties; *ii) measures collection:* the specification of the collecting and adapting process, which allows communicating in what order data collection from data sources is made, the packing of measures into streams and their organization in buffers; *iii) measures processing:* the formalization of the correction and analysis process, which allows basically taking snapshots from the buffer, managing its resources, and processing the time windows. On these windows, several statistical analyses are applied; and *iv) decision making:* the formalization of the decision-making process which allows creating, updating and applying the incremental classifiers for the measures on the entity under analysis, incorporating also a predictive behavior, with the aim of preventing risks. Finally, for illustrating the SDSPbMM approach, we use excerpts of an outpatient monitoring system widely documented in [9].

Following this introduction, Section 2 and 3 summarize the C-INCAMI conceptual framework and the SDSPbMM approach respectively, regarding their components. Then, Section 4 specifies the SDSPbMM main processes, using the SPEM language. Section 5 analyzes the contributions of this research regarding related work. And, finally, the conclusions and future works are summarized in Section 6.

## 2 C-INCAMI Overview

C-INCAMI is a conceptual framework [3,4], which defines the concepts and their related components for the M&E area in software organizations. It provides a domain (ontological) model defining all the terms, properties, and relationships needed to design and implement M&E processes. It is an approach in which the requirements specification, M&E, and analysis of results are performed for satisfying a specific information need in a given context. In C-INCAMI, concepts and relationships are meant to be used along all the M&E activities. This way, a common understanding of data and metadata is shared among projects fostering more consistent analysis.

C-INCAMI is structured in six components, namely: i) *M&E project*, ii) *Nonfunctional Requirements*, iii) *Context*, iv) *Measurement*, v) *Evaluation*, and vi) *Analysis and Recommendation*.

The *M&E project definition* component (not shown in Fig. 1), defines and relates a set of project terms needed to deal with M&E activities, methods, roles and artifacts.

The *Nonfunctional requirements* component (*requirements* package in Fig. 1) allows specifying the *Information Need* of any M&E project. The information need identifies the *purpose* (e.g. "understand", "predict", "monitor", etc.) and the *user viewpoint* (e.g. "patient", "final user", etc); in turn, it focuses on a *Calculable Concept* –e.g. system quality, quality of vital signs- and specifies the *Entity Category* to evaluate –e.g. a resource, system, etc. A calculable concept can be defined as an abstract relationship between attributes of an entity and a given information need. This can be represented by a *Concept Model* where the leaves of an instantiated model are *Attributes*. Attributes can be measured by metrics.

For the *context* package, one concept is *Context*, which represents the relevant state of the situation of the entity to be assessed with regard to the information need. We consider Context as a special kind of *Entity* in which related relevant entities are involved. To describe the context, attributes of the relevant entities are used –which are also Attributes called *Context Properties* (see [4] for details).

The *Measurement* component, includes the concepts and relationships intended to specify the measurement design and implementation. Regarding measurement design, a *Metric* provides a *Measurement* specification of how to quantify a particular attribute of an entity, using a particular *Method* (i.e. procedure), and how to represent its values, using a particular *Scale*. The properties of the measured values in the scale with regard to the allowed mathematical and statistical operations and analysis are given by the *scaleType*. Two types of metrics are distinguished. *Direct Metric* is that for which values are obtained directly from measuring the corresponding entity's attribute, by using a *Measurement Method*. On the other hand, the *Indirect Metric* value is calculated from other direct metrics' values following a *formula* specification and a particular *Calculation Method*.

For measurement implementation, a *Measurement* specifies the task by using a particular metric description in order to produce a *Measure* value. Other associated metadata is the *data collector name* and the *timestamp* in which the measurement was performed.

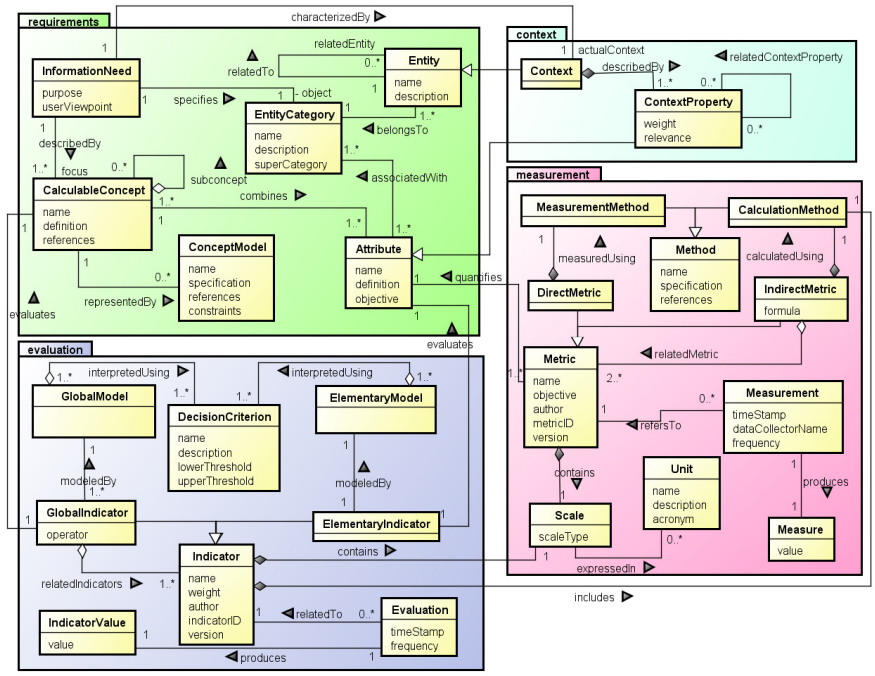


**Fig. 1.** Main components, concepts and relationships for the C-INCAMI conceptual framework

The *Evaluation* component includes the concepts and relationships intended to specify the evaluation design and implementation. It is worthy to mention that the selected metrics are useful for a measurement tasks as long as the selected indicators are useful for an evaluation tasks in order to interpret the stated information need. *Indicator* is the main term, and there are two types of indicators. First, *Elementary Indicator* that evaluates attributes combined in a concept model. Each elementary indicator has an *Elementary Model* that provides a mapping function from the metric's measures (the domain) to the indicator's scale (the range). The new *scale* is interpreted using agreed *decision criteria,* which help analyze the level of satisfaction reached by each elementary nonfunctional requirement, i.e. by each attribute. Second, *Partial/Global Indicator,* which evaluates mid-level and higher-level requirements, i.e. sub-characteristics and characteristics in a concept model. Different aggregation models (*GlobalModel)* can be used to perform evaluations. The global indicator's value ultimately represents the global degree of satisfaction in meeting the stated information need for a given purpose and user viewpoint. As for the implementation, an *Evaluation* represents the task involving a single calculation, following a particular indicator specification –either elementary or global-, producing an *Indicator Value.*

The SDSPbMM strategy reuses the C-INCAMI conceptual base, in order to obtain a repeatable and consistent data stream processing. Thus, streams are basically measures (i.e., raw data usually coming from sensors), which are linked accordingly with

the metadata based on C-INCAMI, such as the entity being measured, the attribute and its corresponding metric with the scale and measurement/calculation procedure, the trace group, among others. For a given data stream, not only measures associated to metrics of attributes are tagged but also measures associated to contextual properties as well. Thanks to each M&E project specification is based on C-INCAMI, the processing of tagged data streams are then in alignment with the project objective and information need, allowing thus traceability and consistency by supporting a clear separation of concerns. For instance, for a given project –more than one project can be running at the same time-, it is easy to identify whether a measure is coming from an attribute or from a contextual property, and also its associated scale type and unit. Therefore, the statistical analysis is benefited because the verification for consistency of each measure against its formal (metric) definition can be performed.

In a previous work [2,9], a practical case using C-INCAMI as part of the SDSPbMM approach was illustrated, particularly, for the outpatient monitoring system in the healthcare domain. The aim is that doctors of a healthcare centre could avoid adverse reactions and major damage in the health of outpatients, if they had a continuous monitoring over them. That is, doctors should have a mechanism by which can be informed about unexpected variations and/or inconsistencies in health indicators defined by them, as experts. Therefore, there exists some proactive mechanism based on health metrics and indicators that produces an on-line report (alarm) for each risky situation associated to the outpatient under monitoring. So, the information need is "*to monitor the principal vital signs of an outpatient when he/she is given the medical discharge from the healthcare centre*". The entity under analysis is the *outpatient*. According to medical experts, the *corporal temperature*, the *systolic arterial pressure*, the *diastolic arterial pressure* and the *cardiac frequency* represent the relevant attributes of the outpatient vital signs to be monitored. They also consider as necessary monitoring contextual properties such as the *environmental temperature, environmental pressure, humidity,* and *patient position* (i.e. latitude and longitude). In addition, the quantification of attributes and contextual properties is performed by well-defined metrics. The definition of the information need, the entity, its associated attributes, metrics, etc. are part of the above C-INCAMI components (see more details in [3,4]).

## 3 SDSPbMM Overview

The SDSPbMM [2,9] strategy proposes a flexible approach in which cooperative components and processes are specialized for data stream management with the main goal of fostering proactive decision making. In this sense, SDSPbMM allows the automation of data collection and adaptation processes supporting also the incorporation of heterogeneous data sources; the correction and analysis processes supporting the early detection of problems typical of data such as missing values, outliers, etc.; and online decision-making processes based on formal definitions of current/updated classifiers. Fig 2 depicts a schema of the SDSPbMM approach.

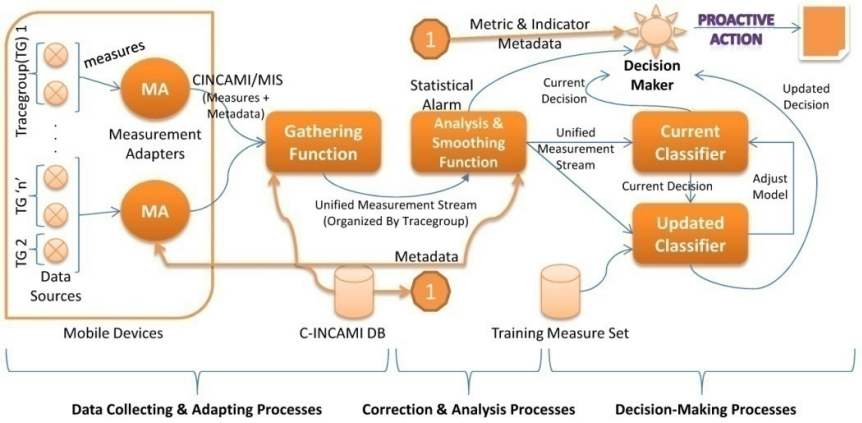


**Fig. 2.** Schema for the Strategy for Data Stream Processing based on Measurement Metadata

In the nutshell, the measurement stream is informed by each heterogeneous data source to the *Measurement Adapter* (MA). The MA incorporates the metadata (e.g. metric ID, context property ID, etc.) associated to each data source into the stream in order to transmit measures to the *Gathering Function* (GF). Such measures are organized in GF by their metadata and then sent to the *Analysis & Smoothing Function* (ASF). ASF performs a set of statistical analysis on the stream in order to detect deviations or problems with data, considering its formal definition (as per C-INCAMI DB). In turn, the incremental classifiers (i.e. the *Current and Updated Classifiers*) analyze the arriving measures and act accordingly triggering alarms in case a risk situation arises.

SDSPbMM is made up of the following components/processes: *Data Collection and Adaptation*; *Correction and Analysis*; and *Decision Making*, which are summarized below (and Section 4 shows these from the process specification standpoint):

### 3.1 Data Collection and Adaptation

Data collection and adaptation deal with how to adapt different measurement devices to collect measures and then communicate them to correction and analysis processes. The main components (see Fig. 2) are data sources, measurement adapters and the gathering function. In short, measures are generated from heterogeneous data sources, and sent continuously to the MA. MA can usually be embedded in mobile devices, but also can be embedded in any computing device associated to data sources. It incorporates the measured values join to the M&E project metadata respectively, sending in turn both to the GF. In turn, GF introduces streams into a buffer organized by trace groups –a flexible way to group data sources established dynamically by the M&E project director. This organization allows consistent statistical analysis at trace group level, without representing additional processing loads. Within each trace group, the organization of measures is tracked by metric. This fosters a

consistent analysis among different attributes (e.g. corporal temperature, cardiac frequency, etc.), which are monitored by a given trace group for a particular outpatient. Also, homogeneous comparisons of attributes can be made for different trace groups (or outpatients). Moreover, GF incorporates *load shedding techniques* [11], which allow managing the queue of services associated to measures, thus mitigating overflow risks regardless of how they are grouped.

### 3.2 Correction and Analysis Component

Correction is based on statistical techniques where data and their associated metadata allow richer (semantic) analysis. The semantic lies in the formal definition of each M&E project regarding the above C-INCAMI M&E framework. It is important to remark that the formal definition of each project is made by experts. In this way, such a definition becomes a reference pattern in order to determine if a particular measure value is meaningful and consistent with regard to its associated metric specification. Once the measures are organized in the buffer, the component applies *descriptive, correlation* and *principal components analysis*. These techniques allow detecting inconsistent situations, trends, correlations, and/or identifying system components that incorporate more variability. If some situation is detected in ASF (see Fig. 2), a statistical alarm is triggered to the *decision maker* (*DM*) element in order to evaluate whether it is necessary to send an external alarm (via e-mail, SMS, etc.) for reporting on this situation to medical staff.

### 3.3 Decision-Making Component

Once the statistical analysis was performed, the unified streams are communicated to the *current classifier* (CC) component, which classifies measures to decide whether they correspond to a risky situation and to accordingly inform such decision to DM. Simultaneously, CC is regenerated by incorporating the unified streams to the training measure set, and then producing a new model named *Updated Classifier* (UC) in Fig. 2. Later, the UC classifies the unified streams and produces an updated decision notifying to DM. Ultimately, DM evaluates if both decisions (from CC and UC) correspond to a risk situation and its probability of occurrence. Finally, regardless the selected decision made by DM, the UC becomes the CC replacing the previous one, only if an improvement in the classification capacity according to the adjustment model based on ROC (*Receiver Operating Characteristic*) [12] curves exists.

### 3.4 Contribution of Metadata to Measures

In this subsection the added value of metadata for data interoperability, consistency and computability is addressed. As mentioned in subsection 3.1, measures are sent from heterogeneous data sources to the GF component through MA. When MA receives data streams from each data source, it incorporates metadata accordingly to a common stream –independently that measures come from several data sources-, and transmit it by means of the C-INCAMI/MIS (*Measurement Interchange Schema*) [9]

schema to the GF component. Thus, before sending measures each data source must configure just once each metric that quantifies each attribute (e.g. the cardiac frequency attribute) of the target entity (e.g. an outpatient), and the included contextual properties (e.g. environmental temperature). This allows to the MA be aware of how such metadata might be embedded into the stream.

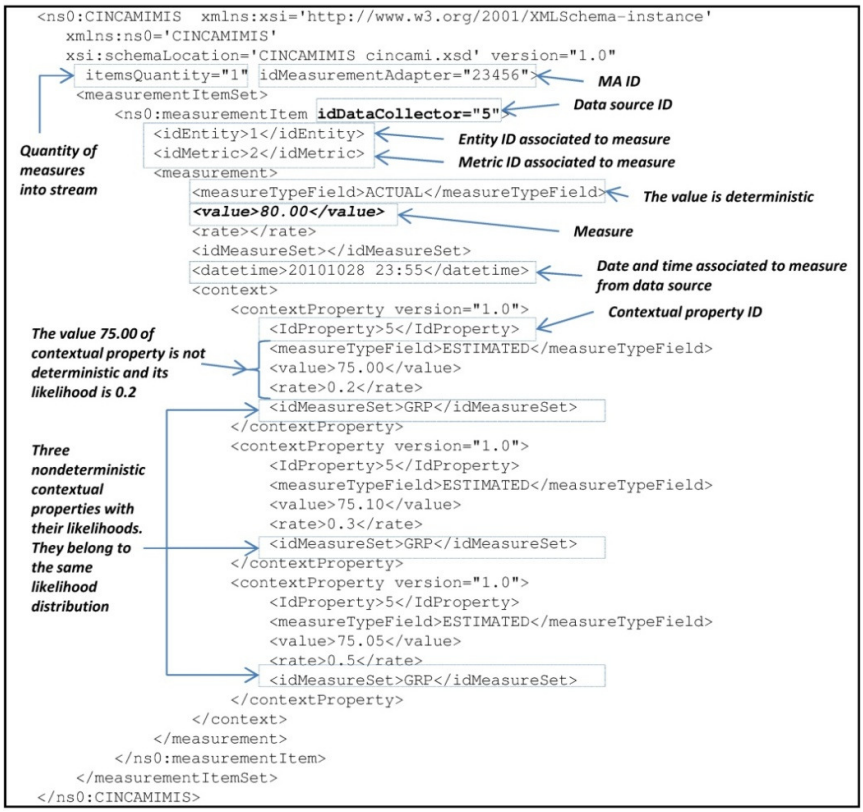


**Fig. 3.** Annotated XML schema of a C-INCAMI/MIS stream.

Hence, CINCAMI/MIS is an XML (*eXtended Markup Language*) schema –based on the C-INCAMI conceptual base as discussed in Section 2-, which copes with interoperability issues in the provision of data from heterogeneous devices, and their further organization.

In Fig. 3 an annotated schema of a C-INCAMI/MIS stream is depicted. For each sent stream, MA incorporates to the raw data –e.g. the value 80– the structure of C-INCAMI/MIS schema, indicating the correspondence of each measure with each attribute and contextual property. For instance, in the message of Fig. 3, IDEntity=1 represents the *outpatient* entity, IDMetric=2 the metric *value of cardiac frequency,* and IDProperty=5 the metric value of *environmental humidity percentage* with regard

to the outpatient location –representing thus a contextual property. Therefore, the metadata in the message clearly includes a set of information which allows keeping a link between a measure value and the origin of data to identify the data source, the metric and entity ID, among others. This information allows increasing the consistency in the processing model for each M&E project definition.

Let's suppose, for example, a value of 80 associated to a cardiac frequency of an outpatient is arrived. Then, the following basic questions can be raised: What does it represent? Which unit of measure does it have? Which mathematical and statistical properties have the value regarding the scale type? Is it good or bad? What is good and what is bad, i.e. what are the decision criteria of the indicator? Could any software process the measure?

Ultimately, if the stream metadata were not available, many questions as those raised above could not be answered in a consistent way. Hence, the correct computability of measures can be hampered and the analysis can be skewed.

## 4 Specification for the SDSPbMM Processes

### 4.1 General Process View

SDSPbMM includes three core processes and one support process, as shown in Fig. 4. On the one hand, the three core processes are related to: a) the collecting and adapting of measures; b) The correction and analysis of the collected measures; and c) the decision-making process based on measures and indicator values. On the other hand, the support process is associated with the configuration of data sources with regard to a particular M&E project.

The three core processes depend on the configuration process (see Fig. 4). From a particular M&E project, the configuration process establishes the correspondence among data sources associated with a MA and the metrics associated with attributes of the target entity and its context properties.

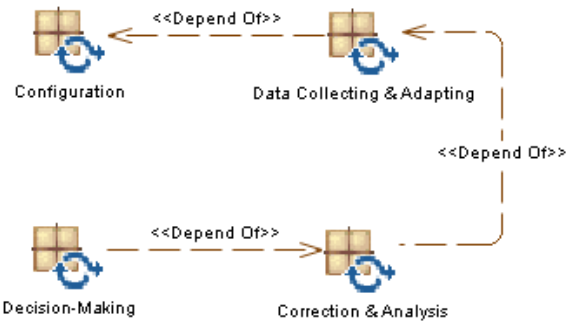


**Fig. 4.** Dependency between processes in the SDSPbMM strategy

Next, the collecting and adapting process collects the measures from the data sources, and then it packs into the stream, the measured values together with the M&E project metadata accordingly. The stream is organized in a MA in base to a measurement interchange schema, and then is introduced in a central buffer. This organization schema is known as C-INCAMI/MIS, as commented in subsection 3.4.

As data streams arrive to the central buffer, the correction and analysis process will take snapshots from the buffer and then will clean and release its resources, considering a processing time window. On these windows, several statistical analyses will be applied and the results will be compared with the formal definition of the M&E project. In case of deviations, these are notified as alarms to the decision-making process.

The objective of the decision-making process is to prevent risky situations as described in sub-section 3.3. Next, we specify each process from the functional and behavioral process modeling standpoint.

### 4.2 Configuration Process

Almost all processes rely to a great extent on the *Configuration* process since it establishes the correspondence between the measurement physical devices -represented by data sources in Fig. 2-, and the attributes/metrics associated to a target entity and its context properties.

Initially, the process reads the local configuration from the MA (see Fig. 5.b), which allows to know whether a previous configuration exists which could be updated, or if there is a new configuration. Next, the process verifies the server availability allowing to get access to the formal definition of each active M&E project structured as per C-INCAMI metadata. Thus, for the selected project, the process configures its data sources. Also, the metrics associated with both attributes of the target entity and context properties can be set up.

For instance, if we want to configure a metric associated with a target entity, the process will request (by using web services) the defined entities related to the selected project. Once that target entity was chosen, the process will ask for the attributes associated with the entity and then will choose one of them. With the selected attribute, the process will ask for the metrics defined for the given attribute and then will choose one of them. Finally, the metric and the data source can be linked if and only if the required precision by the selected metric is satisfied by the data source device.

Similarly, if we want to configure a metric associated with a context property, the process will get the contexts defined for the selected M&E project and then will choose one of them. Next, the process will get the context properties associated with the selected context and then will choose one of them. Then, the workflow followed is like that described above. Fig. 5.a) and b) show that both configurations are locally updated allowing as well to continue configuring those elements. Note also that more than an active M&E project can be configured; additionally, a MA could be simultaneously sending measures to several M&E projects.

Finally, the configuration process is performed just once at the MA start-up, as commented above.

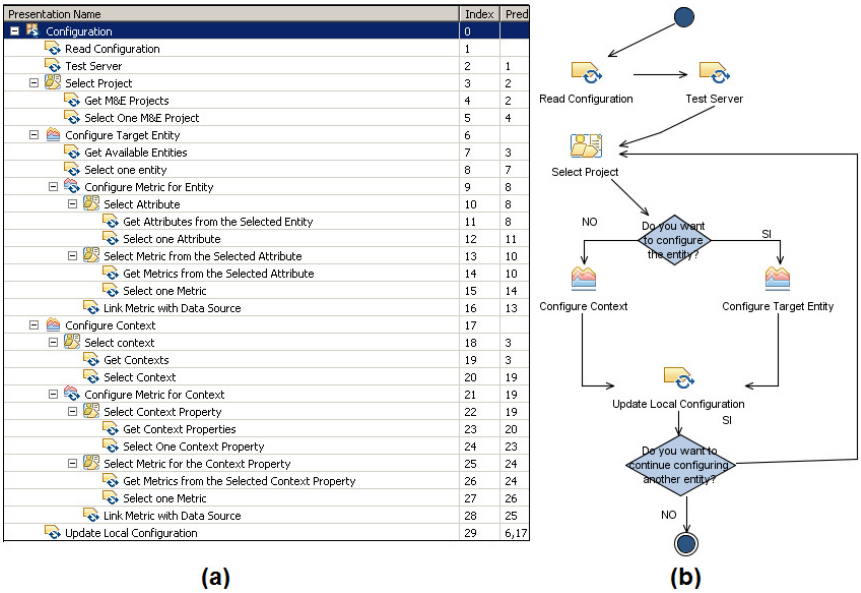


**Fig. 5.** (a) WBS for the *Configuration* process, (b) Activity diagram for this process

### 4.3 Data Collecting & Adapting Process

Once data sources associated with a MA have been configured with regard to one or more M&E projects through the above configuration process, we know exactly what each metric represents (e.g. with deterministic measure values or not), and if it is related to an attribute of a target entity or its defined context.

Through the collecting and adapting process the measures will be collected from data sources. The measures and the metadata will be jointly packed in an XML stream under the C-INCAMI/MIS schema. The stream is organized in a MA and sent to the central buffer for performing a holistic analysis. Each MA contains a local buffer.

Once the local buffer is initialized as shown in Fig. 6, the workload of the MA is parallelized in: a) the buffer monitoring, and b) The measure collecting, adapting and transmitting. The buffer monitoring assesses the state of usage of its resources, e.g. releasing the memory spaces of the stream sent and compressing the measures with the aim of optimizing the memory resource, considering also that a MA can usually run on mobile devices.

Basically, the collecting activity starts when identifies the active data sources associated with the MA, that is, for those data sources that have answered to the alive request, like a 'ping' in networking. The MA periodically requests data sources for the availability of measures, placing them inside the local buffer. Measures are organized in the buffer by the associated metric. It is worth mentioning that thanks to the configuration process, the metric semantic (metadata) associated with the attribute or context property is beforehand known as well as the associated active M&E project.

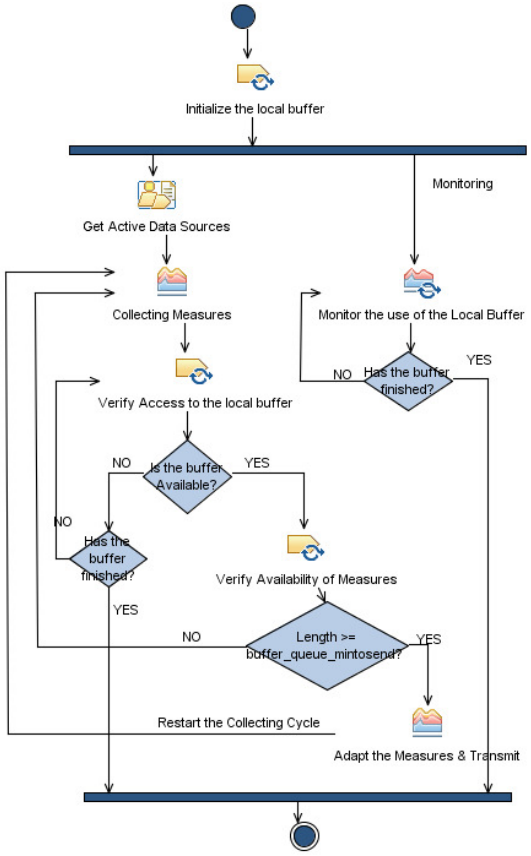


**Fig. 6.** Activity diagram for the *Data Collecting and Adapting* process

When the measures are being incorporated in the local buffer, it is necessary to verify whether the local buffer is available for queries or is in maintenance state at that moment (due to for example to a memory compression task). If the local buffer is ready, the availability of measures for each metric managed by the MA is checked. If a given metric has an amount of measures equal or greater than the *buffer_queue_mintosend* parameter, then the adaptation starts and the measures are sent to the local buffer (see *Gathering Function* in Fig. 2); otherwise, the collection task continues. In this way, the aim of the *buffer_queue_mintosend* parameter is to establish a minimum threshold value which regulates the sending of measures in order to optimize the communication starting time regarding the data volume to be transmitted. When the transmission happened, a snapshot from the local buffer is performed. Then a C-INCAMI/MIS stream is generated incorporating data and metadata accordingly. The stream is sent and incorporated to the central buffer, therefore releasing the local buffer to increment the resource availability. While the local buffer is not

finished, the collecting and sending process continues; but, when the finalization signal is received the application resources are freed.

### 4.4 Correction and Analysis Process

In the correction and analysis process arrived streams to the central buffer are statistically analyzed and contrasted with their formal definition in the M&E project.

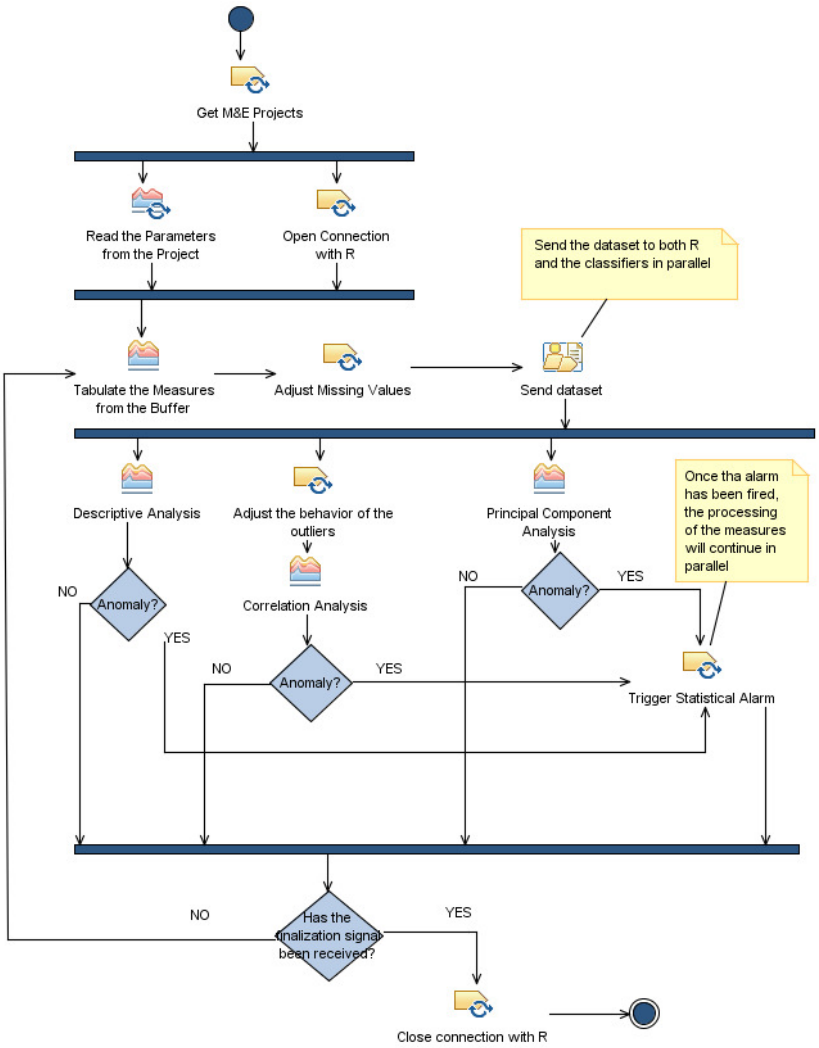


**Fig. 7.** Activity diagram for the *Correction and Analysis* process

At the beginning, the definition of the M&E projects associated with data streams in the central buffer are obtained (see Fig. 7). Next, the parameters defined for each statistical analysis are read (e.g. level of confidence in the correlation analysis for calculating the confidence interval associated with the 'r' coefficient), and the connection with R [13] is opened. Then, a snapshot in a tabular way from the buffer is taken including both metrics of attributes and context properties. The processing of missing values to the dataset is applied, and then sent in parallel to R and classifiers. So, at this point, three different analyses are concurrently run, namely: a) the descriptive analysis; b) the correlation analysis; and, c) the principal component analysis.

Therefore, if in any analysis an anomaly is detected, an alarm is triggered which will be processed by the decision-making process. If there does not exist a finalization signal when the analysis ends, another snapshot from the central buffer is taken and the analysis cycle is restarted. But, if the finalization signal is received, the R connections are closed considering that the current analysis must end before closing connections. For the correlation analysis, it is possible to adjust the behavior of outliers with the aim to minimize their incidence in the analysis. This is important since an outlier could generate a false positive in an association between metrics.

### 4.5 Decision-Making Process

When the decision-making process starts (see Fig. 8), on the one side, the data sent from the tabular processing window (generated in the previous process) is read with the end of creating, updating and applying the incremental classifiers for the measures on the entity under analysis, incorporating also a predictive behavior. On the other side, each statistical alarm is analyzed to conclude about its relevance, therefore, incorporating a detective analysis based on the metric semantic (metadata).

This process checks continuously whether pending statistical alarms exist for their analysis. If there exist alarms, these are individually analyzed because in the M&E project definition is possible that two metrics can be correlated. For example, an indirect metric is calculated from other direct metrics, so there is a possibly of correlation among them. Such a correlation will be disregarded by the SDSPbMM indicating that represents a false positive alarm, due to the metric calculation method is part of the M&E project definition and does not mean a measured fact from the target entity at hand. Conversely, when the measures come from facts, i.e. the measurement of the target entity attributes and context properties, these are conveyed accordingly.

Once the received data through the tabular processing window is read, if it does not exist a classifier, it will be created using the *Adaptive-Size Hoeffding Tree* (*ASHT*) [14,15] algorithm. Then the classifier integrity is checked with the end of its applicability. At this point, the work is forked into two activities, namely: i) The actual classifier on the new measures is applied in order to get a classification at the moment 't'; and, ii) The actual classifier is updated with the new measures and, then, it is applied on the measures to get a classification at the moment 't+1'. Hence, the ROC curve for both classifiers is calculated [12], which allows determining what classifier will continue as the actual classifier. Later, the two obtained classifications are analyzed in order to verify if a risky situation is associated with them. If a situation of risk is de-

tected, an external alarm for its distribution is sent and the analysis continues; conversely, the analysis simply continues.

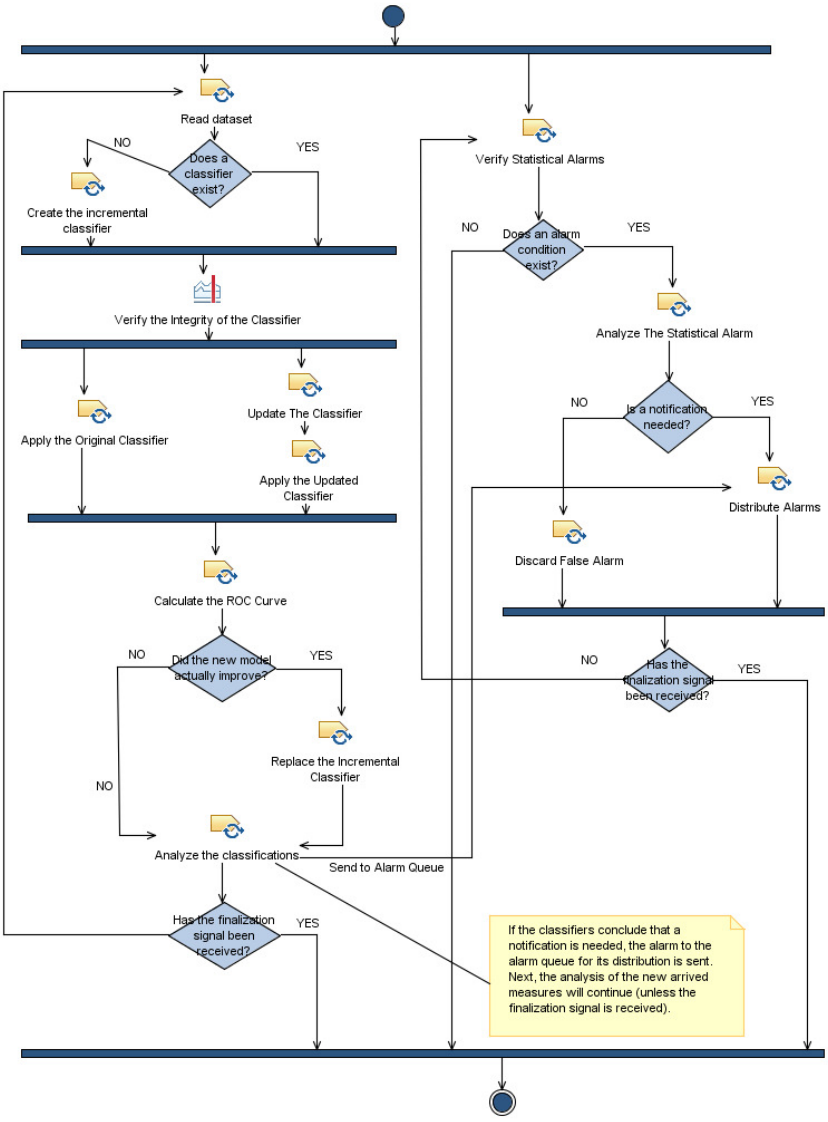


**Fig. 8.** Activity diagram for the *Decision-Making* process

The process ends when the finalization signal is received (see diagram in Fig. 8). However, notice that the finalization signal cannot interfere with the alarm analysis or with the classifier updating; just when both have finished, the process ends.

## 5 Related Work and Discussion

There are many works that are focused on data stream processing from the syntactical point of view. These proposals allow the modeling and querying of data streams in a continuous way based on attributes and their associated values using CQL (*Continuous Query Language)* [16,17]. For example, it has been implemented in several projects such as Aurora & Borealis [7], STREAM [18], and TelegraphCQ [6], amongst others. Our SDSPbMM strategy includes the capability to incorporate metadata of attributes, metrics and indicators based on the C-INCAMI M&E framework, which allow guiding the organization of data streams in the buffer as well as triggering alarms in a proactive way using several statistical analysis as well as based on decisions coming from classifiers. Therefore, more consistent and comparable analyses from a statistical point of view are possible. On the other hand, the SPEM formalization of the SDSPbMM processes also supports well-established specifications, which are repeatable, communicable, and extensible from the different process modeling viewpoints [19].

MavStream [20] is a prototype for a data stream management system, which has the capability of complex-event processing [21] as an intrinsic aspect for data stream management. In this sense, our developed SDSPbMM prototype [9] supports the online data stream analysis with the incorporation of metadata to measures, handling not only measures values (data) coming from attributes of the controlled target entity but also those coming from contextual properties related to the situation of the target entity. Also, the prototype can process measures with non-deterministic values, and perform analysis by trace group, which in practical scenarios such as in the case for monitoring of outpatients (as commented in the end of Section 2, and detailed in [2,9]), represent crucial features. Moreover, it supports the alarm generations in a proactive way with statistical base.

Nile [5] is a data stream management system based on a conceptual framework for detection and tracking of phenomena or situations supported by deterministic measures. Our prototype unlike Nile allows the incorporation of heterogeneous data sources embracing not only deterministic but also nondeterministic measures.

Singh *et al* [22] introduce a system architecture for a formal framework of data mining oriented to the situation presented in [23]. This system is used in medical wireless applications and shows how the architecture can be applied to several medical areas such as diabetes treatment and risk monitoring of heart disease. In our humble opinion, this system neglects central issues for assuring repeatability and interoperability because it lacks a clear specification of metrics both for entity attributes and contextual properties, indicators, scales and scale types, among other metadata.

Lastly, SECRET [24] is a descriptive model which allows to the users to analyze and understand the behavior of stream processing engines (SPE) from window-based queries. This approach takes into account the concern that exists among several SPE proposals (regardless whether they are academic or commercial) related to the processing of semantic diversity. The aim of our strategy is basically different because: i) it deals with the data stream processing, which is driven by metadata of measures, and ii) it is based on a well-defined SPEM workflow.

## 6 Conclusions and Future Work

In this research, we have pointed out how measures in data streams grouped and linked with metadata based on the C-INCAMI M&E framework, allow an efficient organization of measures which foster a sounder consistency and comparability in statistical analysis, since they specify not only the formal metadata of measured attributes but also metadata for context properties. Furthermore, it is possible to perform specific analyses at trace group or at a more general level, comparing values of metrics among different trace groups in order to identify, for example, deviations of measure values against their formal definition, the main system variability factors, as well as relationships among variables.

In addition, we have particularly emphasized the process specification associated with the components of the SDSPbMM strategy. In this way, we have used the SPEM standard language to formalize some process views of SDSPbMM. SPEM supports well-established specifications, which are repeatable, communicable, and extensible. Specifically, we have shown the activity diagrams for the configuration of heterogeneous data sources, the collection and delivery of measures from data sources, the analyses and correction, in addition to the role of statistical analysis and classifiers for decision making. These issues represent the particular contribution of this manuscript since before the SDSPbMM processes i.e., the sequences, parallelisms, feedback loops, among other activity aspects, were not formalized. So the communicability of the strategy was somewhat hindered from the processing workflow viewpoint.

As future work, on the one hand, other strategies oriented to complement the preventive action in the decision-making process will be analyzed. For instance, the idea is to analyze non-supervised mechanisms in which the previous training is not required, as is the case of clustering techniques. Also, the organizational memory like bidirectional and active feedback schema of the incremental classifiers has been proposed in [25]. So the primary aim in this line of research is to complement the non-supervised mechanisms and the organizational memory with the current incremental classification trees of the ASHT type. On the other hand, with the purpose to enrich semantically C-INCAMI M&E terms with process terms, a process conceptual base structured in an ontology was specified in [26]. This work will help us to enhance our data stream processing strategy based on measurement metadata, not only from its process specification but also from its conceptual base point of view.

**Acknowledgments.** Thanks to the support given from the Argentinean Science and Technology Agency in the POIRE 2013-10, PICTO 2011-0277 projects, and in the 066/12, 09-F047 projects at UNLPam, Argentina.

## References

1. Gehrke J., Balakrishan J. & Namit H., Towards a Streaming SQL Standard, Proceedings of the VLDB Endowment, vol. 1, no. 2, pp. 1379-1390, August 2008.
2. Diván M., Olsina L., & Gordillo S., Strategy for Data Stream Processing Based on Measurement Metadata: An Outpatient Monitoring Scenario, Journal of Software Engineering

and Applications, vol. 4, no. 12, pp. 653-665, December 2011.
3. Olsina L., Papa F, & Molina H., How to Measure and Evaluate Web Applications in a Consistent Way, in Ch. 13 in Web Engineering. Springer, 2007, pp. 385–420.
4. Molina H. & Olsina L. Towards the Support of Contextual Information to a Measurement and Evaluation Framework, in QUATIC, IEEE CS Press, Lisboa, Portugal, pp. 154–163, 2007.
5. Aref M., Bose W., Elmagarmid R., Helal A., Kamel A., Mokbel I. & Ali M., NILE-PDT: A Phenomenon Detection and Tracking Framework for Data Stream Management Systems, In proc. of VLDB, Trondheim, Norway, 2005, pp. 1295-1298.
6. Chandrasekaran S., Cooper O., Deshpande A., Franklin M., Hellerstein J., Hong W., Madden S., Reiss F., Shah M. & Krishnamurthy S. TelegraphCQ: An Architectural Status Report, IEEE Data Engineering Bulletin, vol. 26, 2003.
7. Ahmad Y., Balazinska M., Cetintemel U., Cherniack M.. Hwang J., Lindner W., Maskey A., Rasin A., Ryvkina E., Tatbul N., Xing Y., Zdonik S. & Abadi D. The Design of the Borealis Stream Processing Engine, In proc. of Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, pp. 277-289, 2005.
8. Diván, M., Olsina, L. & Gordillo, S. Procesamiento de Flujos de Datos Enriquecidos con Metadatos de Mediciones, In proc. of XVI Conferencia Iberoamericana en Software Engineering (CIbSE'11), Río de Janeiro, Brasil, pp. 251-257, 2011.
9. Diván, M. Enfoque Integrado de Procesamiento de Flujos de Datos centrado en Metadatos de Mediciones, UNLP, La Plata, PhD Thesis, 2011.
10. SPEM, Software Process Engineering Meta-Model Specification, Object Management Group (OMG), Ver.2.0, 2008.
11. Rundensteiner W., Mani M. & Wei M. Utility-driven Load Shedding for XML Stream Processing, In proc. of International World Wide Web, Beijing, China, pp. 855-864, 2008.
12. Duin R., Tortorella F. & Marrocco C. Maximizing the area under the ROC curve by pairwise feature combination, ACM Pattern Recognition, pp. 1961-1974, 2008.
13. R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria: The R Foundation for Statistical Computing, 2012
14. Bifet A., Holmes G., Pfahringer B., Kirkby R., & Gavaldà R., New Ensemble Methods For Evolving Data Streams, In proc. of ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD). International Conference on Knowledge Discovery and Data Mining, Paris (France), 2009, pp. 139-148.
15. Bifet A., Holmes G., Kirkby R., & Pfahringer B., MOA: Massive Online Analysis, Journal of Machine Learning Research, vol. XI, pp. 1601-1604, 2010.
16. Widom J. & Babu S. Continuous Queries over Data Streams, ACM SIGMOD Record, pp. 109-120, 2001.
17. Bockermann C. & Blom H., Processing Data Streams with The RapidMiner Streams Plugin, Technical University of Dortmund, Dortmund, Germany, Report 2012.
18. The Stream Group. "STREAM: The Standford Stream Data Manager", 2003.
19. Becker, P., Lew, P., Olsina, L.: Specifying Process Views for a Measurement, Evaluation and Improvement Strategy. Journal of Advances in Software Engineering, Software Quality Assurance Methodologies and Techniques, Vol. 2012, pp. 1-27. 2012.
20. Jiang Q. & Chakravarthy S. Stream Data Processing: A Quality of Service Perspective. Springer, 2009.
21. Cugola G. & Margara A., Processing flows of information: From data stream to complex event processing, Journal of ACM Computing Surveys, vol. 44, no. 3, Article No. 15, June 2012.

22. Singh, S., Vajirkar, P. & Lee, Y. Context-aware data mining framework for wireless medical application. Lectures Notes in Computer Science of Springer, vol. 2736, pp. 381-391. 2003.
23. Singh, S., Vajirkar, P. & Lee, Y. Context-Based Data Mining using Ontologies. In Lecture Notes in Computer Science, vol. 2813, pp. 405-418. 2003.
24. Botan, I., Derakhshan, R., Dindar, N., Haas, L., Miller, R. & Tatbul, N., SECRET: a model for analysis of the execution semantics of stream processing systems, In proc. of VLDB Endowment, vol. 3, no. 1-2, pp. 232-243, September 2010.
25. Diván, M., Martín, M. & Olsina, L. Towards the feedback of the Data Stream Processing based on Organizational Memory (in Spanish), in Congreso Nacional de Ingeniería Informática/Sistemas de Información, Córdoba, Argentina, 2013
26. Becker, P., Papa, F., Olsina, L.: Process Conceptual Base for Enriching a Measurement and Evaluation Ontology, In proc. of XVII Conferencia Iberoamericana en Software Engineering (CIbSE'14), Pucón, Chile, pp. 53-66. 2014.