Understanding Data Heterogeneity in the Context of Cyber-Physical Systems Integration

Václav Jirkovský, *Member, IEEE,* Marek Obitko, *Member, IEEE,* and Vladimír Mařík, *Senior Member, IEEE*

***Abstract*—The current gradual adoption of the Industry 4.0 is the research trend that includes more intensive utilization of Cyber-Physical Systems. The computerization of manufacturing will bring many advantages but it is needed to face the heterogeneity problem during an integration of various CPSs for enabling this progress. In this paper we describe various types of heterogeneity with emphasis to a semantic heterogeneity. The Cyber-Physical Systems integration problem is classified into two different challenges. Next, we introduce the approach and the implementation of the semantic heterogeneity reduction with the focus on using Semantic Web technologies for a data integration. Then, the Big Data approach is described for facilitating the implementation. Finally, the possible solution is demonstrated on our proposed Semantic Big Data Historian.**

***Index Terms*—Cyber-Physical System, Data Heterogeneity, Semantic Heterogeneity, Big Data, Industry 4.0, Ontology.**

I. Introduction

NOWADAYS, we are witnessing the research trend in the area of embedded systems which concerns very close integration of computing systems and physical systems, especially with the focus on a control. This trend represents the cornerstone of *Cyber-Physical Systems* (CPSs). CPSs are integrated infrastructures involving communications, computation, control, and sensing. The target of CPSs is a tight integration among controlled physical processes and controlling digital computing systems [1]. CPSs are building blocks of systems which form Industry 4.0 as described in [2] — Smart Grids [3], Smart Cities, Smart Factories, Smart Buildings, and Smart Homes [4]. In general, a CPS consists of three main parts, i.e., a cyber part as a computing core, a physical part as a controlled object (or plant in control terminology), and networks as a communication medium between cyber and physical elements [5]. The CPS architecture is also frequently represented as a two part system consisted of a cyber and a physical part.

This research has been supported by Rockwell Automation Laboratory for Distributed Intelligent Control (RA-DIC), by institutional resources for research by the Czech Technical University in Prague, Czech Republic, and by HydroCon a.s.

Václav Jirkovský is with Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Prague 16000, Czech Republic (e-mail: vaclav.jirkovsky@cvut.cz).

Marek Obitko is with Rockwell Automation R&D Center in Prague, Prague 15500, Czech Republic (e-mail: mobitko@ra.rockwell.com).

Vladimír Mařík is with Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Prague 16000, Czech Republic (e-mail: vladimir.marik@cvut.cz).

Additionally, Cyber-Physical Systems are the essential part for the future IoT. However, CPSs within a complex system are produced by different manufacturers with different data models and interfaces. The heterogeneity among these building blocks represents the important issue to be resolved for successful IoT realization.

A proper understanding of data meaning is necessary for the solution of the heterogeneity problem. The understanding may be achieved by a semantic description of data. Thus, a suitable way for the explicit description of data semantics is provided by ontologies [6]. An expressiveness and versatility of ontologies offer a powerful tool for the heterogeneity reduction process. The ontology utilization can make an integration easier because the typical data integration involves a lot of data mappings and conversions among different data formats. Furthermore, it means that developers have to understand a meaning of data within a given context [7]. Consequently, the current effort in the CPS-based approach will produce many new value-creation opportunities for the future manufacturing. The opportunities include for example multiple systems interacting with each other. This integration critically requires real-time or near to real-time analysis of large amounts of cross-device data. Thus, the nature of CPS data leads to the application of Big Data paradigms [8]. The Big Data paradigms and state of the art technologies can be directly utilized in the context of CPS data. This fact needs to be taken into consideration for CPS data usage and processing.

The motivation of this paper is to offer the complex understanding of the challenging CPSs integration which is essential for success of Industry 4.0. The understanding resides in the localization of integration problems and in the identification of semantic heterogeneity causes together with semantic heterogeneity types. Next, the promising solution for the heterogeneity reduction of CPSs is introduced by means of shared ontology which is demonstrated on the Semantic Big Data Historian. The overall contribution of the paper includes the clarification of the semantic heterogeneity reduction process and facilitation of process usage within a real application.

The paper is organized as follows: first we provide a general overview of Cyber-Physical Systems and their integration challenge. Then, we introduce the data heterogeneity problem, the classification of the heterogeneity kinds, and the main causes of the heterogeneity. After the presentation of appropriate solutions for data heterogeneity reduction suitable

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2016.2596101, IEEE Transactions on Industrial Informatics

for industrial applications we demonstrate the process on data from the hydro-electric power station.

II. CPS: Merging Physical and Virtual Worlds

The term Cyber-Physical Systems (CPSs) refers to a new generation of systems with integrated computational and physical capabilities that can interact with humans through many new modalities. In other words, CPSs are systems with embedded computers as a part of devices, buildings, medical systems, automotive systems, avionics, means of transport, etc. Embedded computers monitor and control the physical processes with loops where physical processes affect computations and vice versa. CPSs are able to:

- capture physical data using sensors;
- affect physical processes using actuators;
- evaluate (and possibly save) captured data, and interact between the physical and virtual world;
- are connected with local and/or global networks — wired and/or wireless.

Furthermore, CPSs are cornerstones of the Industry 4.0 (the 4^{th} Industrial Revolution), mainly of the Internet of Things (IoT) and the Internet of Services (IoS) [9]. Because of Industry 4.0 needs, it is expected that a CPS will utilize the benefits of Cloud Computing, i.e., the cyber-part of the CPS will be provided both on-device and in-cloud [10].

In the context of Industry 4.0, CPSs are interconnected and form more complex systems — the IoT, Smart Factories, and Smart Cities. These complex systems exhibit the heterogeneity problem which is one of the biggest obstacles for a cooperation and so it will be discussed in the following sections.

A. Challenging CPS Integration

CPSs are typically integrated into a more complex system for an improvement of their capabilities. Every system maintains specific data model which is derived from the nature of corresponding physical process or processes. CPSs provide data via an interface to other CPSs or other systems. Reversely, it consumes data from surrounding systems for an enhancement of physical process control. Furthermore, CPSs can share joint data storage — local/distributed/in cloud.

The integration problem can be divided into two distinct problems corresponding to various perspective:

1) *Low-level integration* — interconnections among components of a CPS. The low-level integration is illustrated in the Fig. 1. The CPS consists of a physical part and a cyber part. The physical part involves the physical process and physical objects which provide possibility for process controlling. The cyber part can be divided for clarity into two layers — the first layer (*Platform Layer*) represents a system integration of different devices from various manufacturers and the second layer (*Computational Layer*) represents the computational process which is able to control the physical process according to an implemented logic. In other words, the physical process is modeled first with a physical layer abstraction. Then, the corresponding control system is implemented using a computational (software) layer abstraction. Finally, the control system is deployed on the computation platform modeled with the platform layer abstraction. The different abstraction layers use typically non-compatible semantics, which is the cause of semantic heterogeneity.
2) *High-level integration* — interconnections of various CPSs to form for example IoT. The integration is depicted in the Fig. 2 and will be discussed in detail in the following paragraphs.

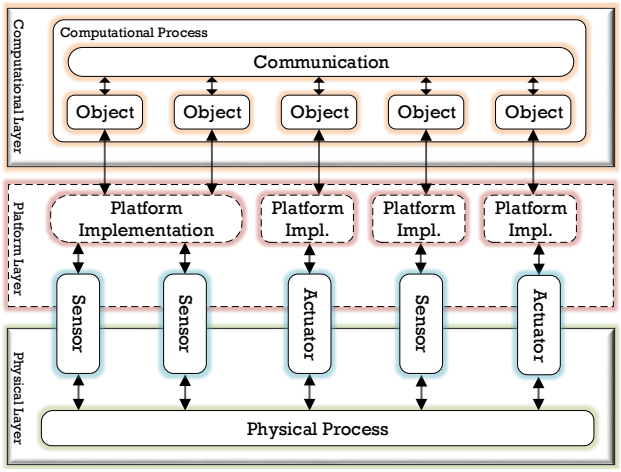


Fig. 1. Cyber-Physical System low-level integration

The primal intention of this paper is to describe the high-level integration for facilitating more complex system creation but the main characteristics of the problem are mostly identical also to the low-level integration.

Nowadays, a common integration of system components relies on *ad hoc* solutions. These solutions can provide very effective systems. However, they may bring many drawbacks — difficult system maintenance, malfunction corrections, adding or adjusting components, re-usability, etc. The example of the high level integration is presented by Smart Living[1] with *Personal IoT* solution. The combination of components consists in the gateway called ATT IOT. A user has to write a script to automate connected components via the Application Programming Interface (API). This approach solves a platform heterogeneity but does not provide any information about a data meaning. Oppositely, the low-level integration could be described as traditional control system development. The original equipment manufacturers (OEMs) purchase system components from suppliers and integrate them into a product. "Since different suppliers implement their products with different strategies, engineering processes, and tools, OEMs always face the integration challenge." [11]

The integration task consists in the unification of CPS interfaces as well as the unification of corresponding data models. This integration becomes more difficult in the case of increasing complexity of systems. There are two dimensions of CPSs integration:

[1]Smart Living - http://smartliving.io

1) *System integration* — the first step of CPSs integration lies in the platform unification because of different interfaces provided by various manufacturers.
2) *Model integration and semantic integration* — model integration covers the identification of corresponding concepts, relations among concepts and their meaning in given context.

A high-level case of semantic integration is illustrated in the Fig. 2. The global model is the key part of whole integration. The model has to map all particular data models and provide proper relations among data concepts. This data model together with implemented transformations acts as "united interface" for direct user queries or an interaction with other systems in the first case; or acts as transformation component for subsequent data storing in global CPS data storage.

III. Data Heterogeneity

In general, a heterogeneity is a feature of all kinds of systems. This feature is both a welcome and an unwelcome feature [12]. On the one hand, the heterogeneity is welcome because of close relation to the system efficiency — more efficient system is more tailored to the problem. On the other hand, the heterogeneity is considered as unwelcome because it causes significant obstacles for the systems interoperability. This situation brings non-trivial dilemma and we need to find balance between the efficiency and the interoperability.

As mentioned earlier, CPSs are produced by various manufactures with various data models and interfaces. Thus, it is needed to ensure an interoperability among them. This need is achievable by means of heterogeneity reduction.

A. Heterogeneity Classification

There are different classifications of heterogeneity types, e.g., in [13]. In this paper, we adopt following types of heterogeneity according to [14]:

- *Syntactic heterogeneity* — occurs when two data sources are not described in the same knowledge representation formalism (e.g., F-logic and OWL in the case of integration of ontologies).
- *Terminological heterogeneity* — means variations in names when referring to the same entity (e.g., different natural language).
- *Semantic heterogeneity* — occurs when different models are used for the same domain of interest (e.g., utilization of different axioms for defining concepts).
- *Semiotic heterogeneity* — denotes different interpretation of entities by different people.

Let us point out that usually several heterogeneity types occur together.

B. Semantic Heterogeneity Definition

In contrast to the other heterogeneity types, a meaning, a clear definition, and causes of semantic heterogeneity are not well understood yet. We would like to contribute to the proper understanding of the semantic heterogeneity and for that we introduce some possible definitions in this section.

The question is how to define the semantic heterogeneity correctly. This is non-trivial task and there is no unique and generally accepted definition. However, there are already some definitions which may provide certain level of understanding.

- The simplest semantic heterogeneity definition can be derived according to Merriam-Webster dictionary[2] as a quality or a state of being made up of parts that are different — related to the meanings of words and phrases.
- More complex definition can be found in [15] as differences in the meaning and use of data that make it difficult to identify the various relationships that exist between similar or related objects in different components.
- Semantic heterogeneity can be defined as differences in the real world interpretation of context, meaning, and use of data [16].

We can extend these definitions as the following semantic heterogeneity definition:

Semantic heterogeneity denotes differences in modeling of different/equivalent concepts and their explicitation (i.e. how the concepts are expressed).

It is important to remark that the semantic heterogeneity is relative feature which is related to particular counterparts — i.e., when given system is considered to have heterogeneous data model in comparison to given surrounding systems then it can be non-heterogeneous in comparison to other systems.

C. Causes of Semantic Heterogeneity

It is important to be able to understand what causes the semantic heterogeneity. The proper understanding of the causes facilitates subsequent heterogeneity reduction.

The main causes of the semantic heterogeneity among data-sources can be identified as different designer influences in the developing processes. In other words, a design autonomy discussed also in [17] includes following conflicts:

- *Difference in coverage* — occurs when data-source models describe different data regions (possible overlapping) at the same level of detail and from the same perspective.
- *Difference in perspective* — occurs when data-source models describe the same data regions at the same level of detail, but from different perspective.
- *Difference in granularity* — occurs when data-source models describe the same data regions from the same perspective, but at different level of detail.
- *Incompatible design specifications* — occurs when different specifications of schema are used.

It is impossible to solve these differences among data-sources in the real world. Therefore consequences have to be solved instead of the causes.

D. Semantic Proximity and Disparity

Various measures were proposed for measuring a proximity or a disparity of objects (semantic distance of objects). The large list of measures and also ways of combining them can

[2]http://www.merriam-webster.com/dictionary

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2016.2596101, IEEE Transactions on Industrial Informatics

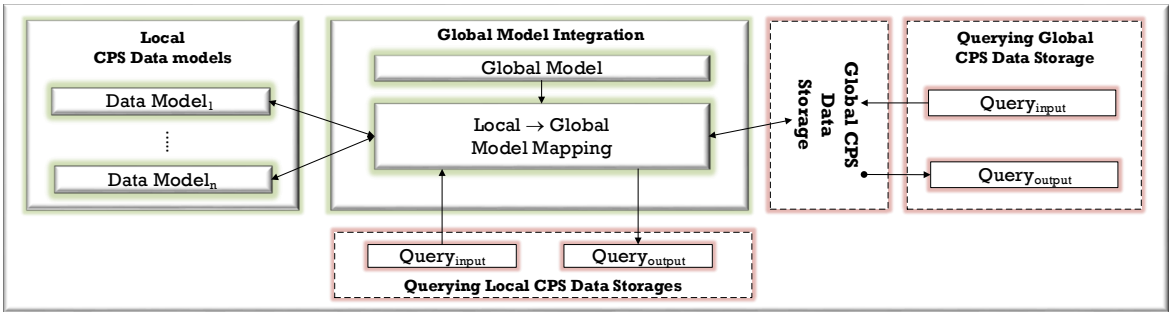


Fig. 2. High-level Cyber-Physical Systems integration schema

be found in for example [14]. Particular measure is not the primary focus of this paper but the general understanding of similarity measure components could be valuable for a reader.

The general semantic distance between two given objects O_1 and O_2 is defined by the 4-tuple given by [18]:

$$d_{sem} = \langle Context, Abstraction, (D_1, D_2), (S_1, S_2) \rangle$$

where D_i is domain of O_i and S_i is state of O_i.

- $Context$ stands for context in which the O_1 and O_2 are being compared. The context may be the same, partly different, or completely different.
- $Abstraction$ denotes the abstraction related to domains of O_1 and O_2.
- (D_1, D_2) identifies the domain definitions related to O_1 and O_2.
- (S_1, S_2) enumerates the states of O_1 and O_2.

Let us point out that usually the measurement of the semantic distance cannot be performed without detailed knowledge about all data and corresponding relations. On the other hand, this measure can provide additional complexity indicator of performed and completed heterogeneity reduction.

E. Why to Deal with Semantic Heterogeneity

An integration of heterogeneous systems means time consuming and exhausting task for domain experts. These experts need deep knowledge about an application area. Especially, when all cooperating parts are developed by different designers then needed coherences among elements may not be obvious.

Inaccuracies in the integration bring out financial loses and within some application can cause important threat (e.g. application in health care, aviation).

Furthermore, the high semantic heterogeneity occurs typically together with stored redundant information. Besides saving used space for data storage when the semantic heterogeneity is resolved it helps to keep data consistency.

F. How to Deal with Semantic Heterogeneity

The easiest way of coping with a data heterogeneity is to take the data model into account during the design of a system. In this case, the data model for the system can be created with the all background knowledge including a knowledge of other systems where the system is anticipated to be integrated — if it is known in advance.

Nevertheless, it is not an exception that a new sensor (or even another system) has to be integrated after the system is designed and when the system is already deployed. Typically, these new parts do not fit into the base system precisely, and therefore new adapters and transformations have to be developed in order to integrate new sensors or systems.

Semantic Heterogeneity Reduction — a process of semantic heterogeneity reduction consists of following steps:

- Identification of heterogeneous elements.
- Potential adjustment of global data model.
- Transformation of heterogeneous data.

The answers to following questions can facilitate the identification of heterogeneous elements:

- Is information (i.e. element) a transformation or a combination of other information?
- How is information used?
- Is similar information related?

A promising solution for the heterogeneity reduction can be found in Semantic Web technologies. The Web is abnormally heterogeneous environment and is based on interconnecting various data sources. The suitable solution can be identified as semantic description of CPSs by means of shared ontology [19] and ontology representation in OWL format [20].

G. Enabling Data Integration of CPSs

In previous paragraphs, the solution based on detailed and proper semantic description of data with the help of OWL was proposed. Unfortunately, the realization is not effortless and not entirely straightforward. In detail, the realization has two problematic aspects — the ontology creation and the system architecture.

First, the creation of the ontology is the cornerstone for achieving a faultless operating system and for potential correct interaction with surrounding systems. However, the creation of the ontology describing many various systems may be exhaustive and therefore the utilization of systems for automatic or semi-automatic ontology creation is appropriate [21].

Next, a subsequent system behavior and system features are influenced by a selection of a way how to handle data. In other words, the relation between the performance and the ease of system management.

The natural approach for data representation and their subsequent handling resides in an ontological model of given data and a triple-store for storing data. The advantages of this

approach is primarily easy management of a data model and no additional effort concerning data storage. The significant drawback is unsatisfactory system performance of many triple-stores [22] for big amount of data. The unsatisfactory performance includes response time of querying and load time depending on number of loaded triples.

The other possible approach is creation of a "hybrid system". The hybrid system consists of the ontological description of data (shared ontology), transformation layer providing interface between ontology and data storage, and data management system (local or distributed) [23]. The main advantage is high performance of data querying. The drawbacks are complicated system management (an adjustment of ontology means changes in the implementation of data transformation) and worse possibility of reasoning in stored data.

Nowadays, the reasons for hybrid system designing are on the decrease because of new promising triple-stores running under Big Data frameworks and RDF[3] extensions coming into existence. On the other hand, the inconvenience of the hybrid approach with the most of traditional data management systems is emphasized by increasing massive data production of CPSs.

H. Big Data and CPSs Semantic Heterogeneity Reduction

The semantic description of data produced by CPSs and expressed in RDF triples causes one significant obstacle. The description is able to carry more information contrary to a database schema but the data volume increases correspondingly. The CPSs systems are able to generate big amount of data and therefore the processing of such data is impossible when classical triple stores are used. For example, we encountered issues regarding loading and querying time and horizontal scalability. This performance problem can be resolved by the Big Data approach.

The term Big Data is used for datasets that are growing so that it becomes difficult to manage them using existing database management concepts and tools [8]. But the term Big Data does not denotes exact amount of data or exact problem. In other words, there is no precise definition.

Therefore the term Big Data is defined by the given data characteristics — also known as "3V definition":

- *Volume* — it is predicted that data will grow 50 times during next 5 years. The systems are moving to processing petabytes and larger amounts of data. It is needed to process the whole produced data to capture all information from data (previously known and hidden as well).
- *Velocity* — the speed at which the data is created, stored, analyzed and visualized. Most of the operations have to be performed in "real" time. For example, an early diagnostics of a failure in production system means a significant financial savings.
- *Variety* — the variety of data for a processing is not only from different systems corresponding with different data schema. Approximately 90% of all new generated data are unstructured. The understanding and right data integration is one of the biggest challenges in data analysis.

[3]http://www.w3.org/TR/PR-rdf-syntax/

TABLE I
NIST – BIG DATA CHARACTERISTICS

Volume	Velocity	Variety	Requires Horizontal Scalability	Relational Limitation	Big Data
No	No	No	No	No	No
No	No	Yes	No	Yes	Type 1
No	Yes	No	Yes	Maybe	Type 2
No	Yes	Yes	Yes	Yes	Type 3
Yes	No	No	Yes	Maybe	Type 2
Yes	No	Yes	Yes	Yes	Type 3
Yes	Yes	No	Yes	Maybe	Type 2
Yes	Yes	Yes	Yes	Yes	Type 3

Are data produced from a CPS or from a conjunction of CPSs suitable for Big Data analysis? It cannot be generalized because the answer is related to a problem and data character. Nevertheless, the Big Data approach fits for many cases.

The National Institute of Standards and Technology (NIST) introduced Big Data characteristics and taxonomy[4] which may solve previously mentioned question. There are three main types of Big Data:

- *Type 1* — represents a problem where non-relational representation is required for effective processing.
- *Type 2* — represents a problem where horizontal scalability is required for effective analysis.
- *Type 3* — represents a problem where non-relational representation as well as horizontal scalability is required for effective analysis.

The table I (introduced by NIST) is derived from the *3V definition* and the Big Data types. The table can answer the question — are the given data of CPSs Big Data?

IV. TESTING SCENARIO

Our proposed shared SHS ontology and Semantic Big Data Historian (SBDH), which is example of semantic CPSs data integration solution, are described in this section together with example application. Generally, a historian software is used in the industrial automation to gather data and then to provide an access and possibly also analytics of measured data and historical data as well. The historian software is usually optimized to fast and compressed storage of data, but not much attention is paid to analytics or heterogeneous data integration.

Data from the hydroelectric power station were chosen for the verification of the proposed SBDH functionality. This example will be more elaborated in the subsection IV-C.

A. Semantic description of Industrial Data

We have proposed a shared ontology (called SHS ontology) for description of industrial data. The cornerstone of our ontology is Semantic Sensor Network (SSN) ontology[5]. The SHS

[4]Michael Cooper and Peter Mell, "Tackling Big Data" slide presentation. NIST Information Technology Laboratory, Computer Security Division, US Department of Commerce, not dated.
http://csrc.nist.gov/groups/SMA/forum/documents/june2012presentations/fcsm_june2012_cooper_mell.pdf

[5]https://www.w3.org/2005/Incubator/ssn/ssnx/ssn

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2016.2596101, IEEE Transactions on Industrial Informatics

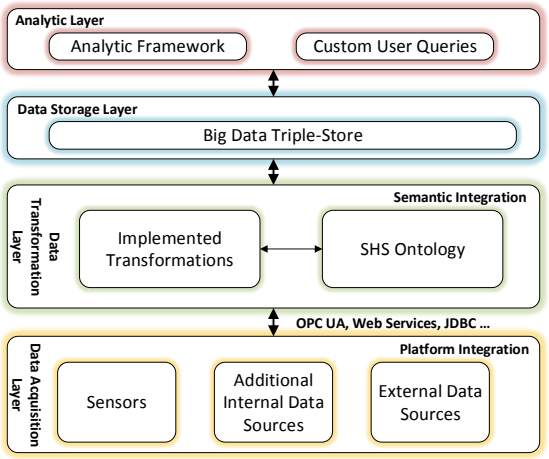


Fig. 3. Architecture of the Semantic Big Data Historian

ontology includes structures for modeling different observations, physical quality, units of measurements, or possibilities of external data sources connections.

Besides the easier management compared to traditional DBMS the semantic data description offers the following advantages:

- Complex querying using SPARQL[6] — for example, what was the temperature trend during afternoons of this week, plus compare it to previous similar weekdays, holidays, after it rained, when different suppliers were used etc.
- Data consistency maintenance by a reasoner.
- Easier understanding of data meaning and context.

B. Semantic Big Data Historian

The architecture of proposed Semantic Big Data Historian, including sensor data gathering, is illustrated in the Fig. 3.

The historian architecture can be divided into four main parts — data acquisition layer, transformation layer, data storage layer, and analytic layer:

- *Data acquisition layer* — data from sensors (e.g., connected via OPC UA [24] connectors), other systems related to given application (e.g., MES/ERP systems, information about shifts), and relevant external data sources (e.g., weather forecast, traffic information). The platform heterogeneity (related to various systems) has to be resolved by this layer.
- *Transformation layer* — transforms data to the unified semantic form according to SHS ontology. This layer corrects damaged data if needed. Next, triples are immediately sent to the triple store (following layer). The semantic heterogeneity is solved by this layer.
- *Data storage layer* — we have evaluated triple stores during our SBDH development. We have identified as the most suitable for our historian software the following ones: 4Store[7], CumulusRDF[8], Hadoop together with Jena Elephas[9].
- *Analytic layer* — this layer provides access to directly connected storage layer for analytics or custom user queries. The following analytic frameworks were chosen as suitable for the historian — KNIME[10], Mahout[11].

Currently the SBDH is in the state of working prototype. Nevertheless, even in this state this system has proved that CPSs may be successfully integrated by means of data representation using shared ontology.

C. Hydroelectric Power Plant Example

Let us describe briefly one of the applications of this approach on data from a hydroelectric power plant. This scenario is used for verifying the concept of CPSs integration by means of SHS ontology and the SBDH. In our application, we process data measured by 38 sensors in the power plant including for example measurement of fall of water, frequency, power factor, and real power. All data from power plant sensors are read with 5 second sampling rate. These data sources are connected via implemented adapters to comply with the SHS ontology.

The significant problem is the performance issue in the context of CPS data processing — especially in the case when data are stored as RDF triples. Our sensors produce $656,640$ samples per day. If we transform these data into triples, then the volume of data is equal to $5,253,120$ triples per day and it corresponds to $1,917$ mil. triples per year.

Moreover, it is needed to involve additional data sources for the improvement of analytic results. We integrated available online information covering meteorological data (temperature and precipitation) and hydrological data (rate of flow, water level and water temperature) for relevant locations. These data consist of actual measured samples as well as the prediction. The sampling rate is 1 hour for meteorological data and 10 minutes for hydrological data. These external data add negligible volume in comparison to data from power plant, but they bring important information. Additionally, we plan the extension of the data by the purchase prices of electricity, by "who is present at power plant premises", etc.

The Fig. 4 represents partially the SHS ontology model of CPSs integration used by the historian. There is the concept model (including concepts from DOLCE ontology[12]) in the upper part and individuals representing one sensor and its measurement for every data source involving CPS in the lower part of the figure.

The collected samples from all data sources (power plant sensors and external sources) are transformed into RDF triples, e.g. *Sensor_RP_0001 hasLocation Generator*. Next, the triples are stored in the corresponding triple store.

As we already mentioned, the demanded analysis includes rapid processing of a huge data volume (e.g., collected during one year). A rapid capturing and an evaluation of information is essential for correct decision making. For these needs the classical triple stores (e.g., Jena TDB) encounter their limits during data upload, processing, and querying. Thus, the storage layer of the SBDH utilizes Hadoop together with Jena

[6]http://www.w3.org/TR/rdf-sparql-query/
[7]http://4store.org
[8]https://code.google.com/p/cumulusrdf/
[9]https://jena.apache.org/
[10]https://www.knime.org
[11]http://mahout.apache.org
[12]http://www.loa.istc.cnr.it/old/DOLCE.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2016.2596101, IEEE Transactions on Industrial Informatics

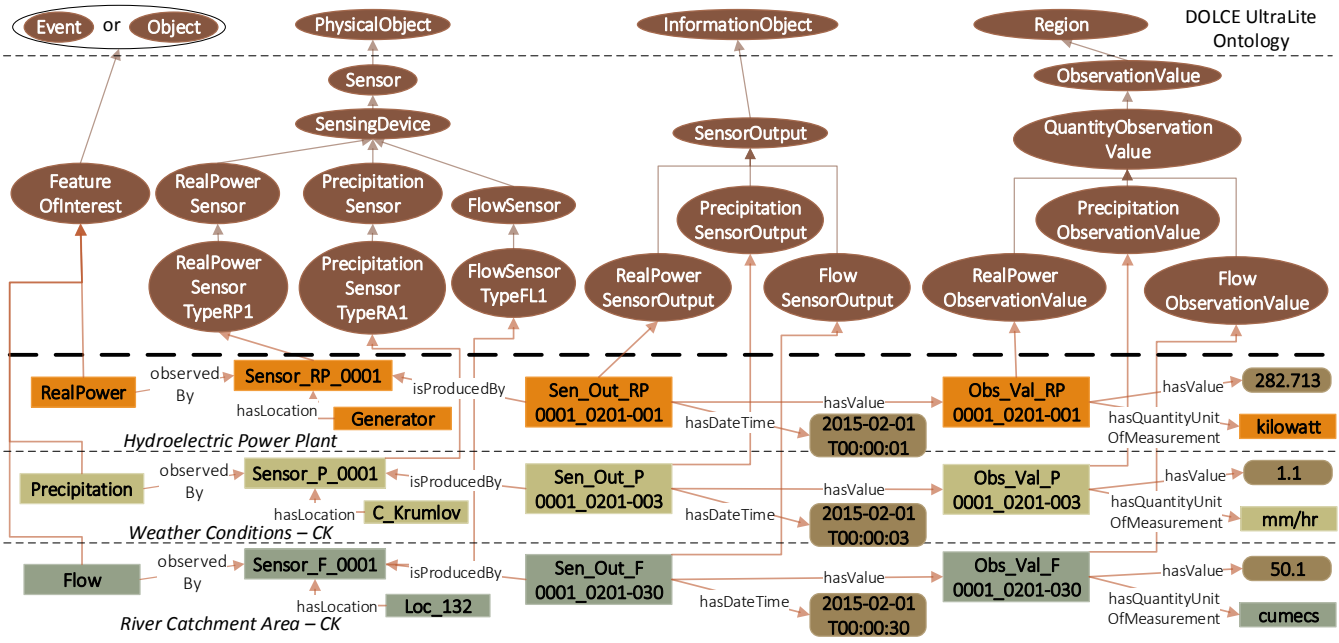


Fig. 4. CPSs integration by means of SHS ontology

Elephas. Data are stored in vertical partitions — RDF triples are partitioned in the Hadoop Distributed File System according to their property value. This platform ensures flexible, robust and scalable environment, which is able to overcome the performance issues of our historian.

The advantage of this way of the integration represented by the SHS ontology and the SBDH are as follows — the ontology is able to describe (with the help of axioms) the reality in its representation. On the contrary, classical schemas (as a database schema) are representation mechanisms that are designed to meet the requirements of a particular application and when the requirements change it is difficult to change the schema and the implementation as well. Next, data can be easily queried in SPARQL. Relationships within data are explicitly described and directly accessible, and therefore SPARQL queries are significantly closer to a user understanding of the problem. For example, the sample query for listing all sensors located in generator is described in Query 1.

Query 1. Listing all sensors in generator

```
PREFIX : <http://www.loa-cnr.it/ontologies/DUL.owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX shs: <http://www.rockwellautomaiton.com/shs#>
PREFIX ssn: <http://purl.oclc.org/NET/ssnx/ssn#>
SELECT ?sensor
        WHERE {?sensor :hasLocation shs:Generator.
        ?sensor a ?sensorType.
        ?sensorType  rdfs:subClassOf+ ssn:Sensor}
```

As we already mentioned, the exploitation of ontology offers possibility to check data consistency and reasoning. Furthermore, the expressivity can be significantly increased by utilization of the Semantic Web Rule Language[13]. The equation 1 illustrates a sample SWRL rule for the malfunction detection caused by unacceptable rotations of the generator.

[13]https://www.w3.org/Submission/SWRL/

$$Generator(?g) \ \wedge \ hasRotation(?g, ?rot) \\ \wedge \ swrlb:greaterThan(?rot, 1500) \ \rightarrow \ hasFailure(?g) \quad (1)$$

V. Discussion and Conclusions

The needs to resolve semantic heterogeneity are pervading many domains, including communication between people and connected systems. This can be viewed as a generalization of the heterogeneity reduction within integrated CPSs.

Unfortunately, CPSs integration is very complex task. It is difficult to understand data from various data sources even within one system. Furthermore, it is suitable to include information from additional systems to uncover unknown dependencies among data, but the subsequent analysis is exhausting task. As mentioned earlier, CPSs are cornerstones of Industry 4.0 and therefore enabling faultless and effective systems integration is necessary for the deployment of Industry 4.0 ideas.

In this paper, we have introduced details about low-level and high-level integration related to CPSs and summarized heterogeneity classification together with semantic heterogeneity definition. We have also presented the general semantic disparity measure and semantic heterogeneity reduction with the help of shared ontology and Big Data approach. Finally, we have introduced and described our SBDH as a demonstration of heterogeneity reduction within a CPS.

The next complex aspect concerning the task of heterogeneity reduction is the establishment of balance between efficiency and interoperability. This is the significant reason why we cannot recommend any generally usable solution so far. However, the solution based on semantic description of data for CPSs integration using shared ontology was introduced and demonstrated by means of the Semantic Big Data Historian prototype. This approach includes many advantages — easy

management of the whole system, wider possibilities for integration, easy understanding of given data model directly from data (additional documentation for manual integration is not needed), data consistency maintenance by a reasoner, etc. On the other hand, the shared ontology utilization requires that all data sources have nearly the same view on a problem and the same level of granularity.

The future work will benefit from the utilization of Semantic Web technologies which offer many interesting options. For example, an addition of "Plug&Play" capability to CPSs represents next step for enabling Industry 4.0 adoption. This capability expects that every CPS includes explicit OWL description of the connected system and the data it provides. If the OWL description reflects the shared ontology then no other configuration is needed. For example, the OPC UA standard offers capability to transfer information in the form of RDF triples in its metadata in the case of smart sensors and thus makes the sensor "Plug&Play" capability possible.

REFERENCES

[1] X. Cao, L. Liu, W. Shen, A. Laha, J. Tang, and Y. Cheng, "Real-time misbehavior detection and mitigation in cyber-physical systems over WLANs," *IEEE Trans. Ind. Informat.*, vol. PP, no. 99, pp. 1–1, 2015.

[2] M. Obitko and V. Jirkovský, "Big data semantics in industry 4.0," in *Industrial Applications of Holonic and Multi-Agent Systems*. Springer International Publishing, 2015, pp. 217–229.

[3] H. Georg, S. Muller, C. Rehtanz, and C. Wietfeld, "Analyzing cyber-physical energy systems: The INSPIRE co-simulation of power and ICT systems using HLA," *IEEE Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2364–2373, Nov 2014.

[4] A. Corrales Paredes, M. Malfaz, and M. Salichs, "Signage system for the navigation of autonomous robots in indoor environments," *IEEE Trans. Ind. Informat.*, vol. 10, no. 1, pp. 680–688, Feb 2014.

[5] K.-J. Park, J. Kim, H. Lim, and Y. Eun, "Robust path diversity for network quality of service in cyber-physical systems," *IEEE Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2204–2215, Nov 2014.

[6] G. Xiao, J. Guo, L. Da Xu, and Z. Gong, "User interoperability with heterogeneous IoT devices through transformation," *IEEE Trans. Ind. Informat.*, vol. 10, no. 2, pp. 1486–1496, May 2014.

[7] W. He and L. Da Xu, "Integration of distributed enterprise applications: A survey," *IEEE Trans. Ind. Informat.*, vol. 10, no. 1, pp. 35–42, Feb 2014.

[8] S. Singh and N. Singh, "Big data analytics," in *Proceedings of ICCICT 2012*, Oct 2012, pp. 1–4.

[9] L. D. Xu, W. He, and S. Li, "Internet of things in industries: A survey," *IEEE Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2233–2243, Nov 2014.

[10] A. W. Colombo, S. Karnouskos, and T. Bangemann, "Towards the next generation of industrial cyber-physical systems," *Industrial Cloud-Based Cyber-Physical Systems*, pp. 1–22, 2014.

[11] J. Sztipanovits, X. Koutsoukos, G. Karsai, N. Kottenstette, P. Antsaklis, V. Gupta, B. Goodwine, J. Baras, and S. Wang, "Toward a science of cyber-physical system integration," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 29–44, Jan 2012.

[12] P. R. Visser, D. M. Jones, T. J. Bench-Capon, and M. J. Shave, "Assessing heterogeneity by classifying ontology mismatches," in *Proceedings of the FOIS*, vol. 98, 1998.

[13] C. H. Goh, "Representing and reasoning about semantic conflicts in heterogeneous information systems," Ph.D. dissertation, Massachusetts Institute of Technology, 1996.

[14] J. Euzenat, P. Shvaiko *et al.*, *Ontology matching*, 2nd ed. Springer, 2013.

[15] J. Hammer and D. McLeod, "An approach to resolving semantic heterogeneity in a federation of autonomous, heterogeneous database systems," *International Journal of Cooperative Information Systems*, vol. 2, no. 1, pp. 51–83, 1993.

[16] D. George, "Understanding structural and semantic heterogeneity in the context of database schema integration," *Journal of the Department of Computing, UCLAN*, vol. 4, pp. 29–44, 2005.

[17] C. Batini, M. Lenzerini, and S. B. Navathe, "A comparative analysis of methodologies for database schema integration," *ACM Comput. Surv.*, vol. 18, no. 4, pp. 323–364, Dec. 1986. [Online]. Available: http://doi.acm.org/10.1145/27633.27634

[18] A. P. Sheth and V. Kashyap, "So far (schematically) yet so near (semantically)," in *Proc. IFIP WG2.6 Database Semantics Conf. Interoperable Database Systems*, vol. DS-5. D. Hsiao, E. Neuhold, and R. Sacks-Davis, eds., 1992, pp. 283–312.

[19] N. F. Noy, "Semantic integration: a survey of ontology-based approaches," *SIGMOD Rec.*, vol. 33, no. 4, pp. 65–70, 2004.

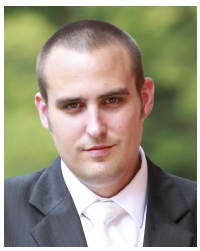
[20] M. Ruta, F. Scioscia, G. Loseto, and E. Di Sciascio, "Semantic-based resource discovery and orchestration in home and building automation: A multi-agent approach," *IEEE Trans. Ind. Informat.*, vol. 10, no. 1, pp. 730–741, Feb 2014.

[21] B. Fortuna, M. Grobelnik, and D. Mladenic, "Ontogen: Semi-automatic ontology editor," in *HCI International 2007*, M. J. "Smith and G. Salvendy, Eds. Springer Berlin Heidelberg, 2007, vol. 4558, pp. 309–318.

[22] K. Rohloff, M. Dean, I. Emmons, D. Ryder, and J. Sumner, "An evaluation of triple-store technologies for large data stores," in *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops*. Springer Berlin Heidelberg, 2007, vol. 4806, pp. 1105–1114.

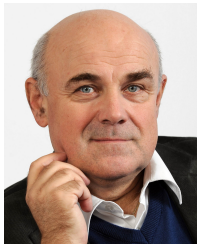
[23] Y. Liang, H. Bao, and H. Liu, "Hybrid ontology integration for distributed system," in *Proceedings of the Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, vol. 1, July 2007, pp. 309–314.

[24] A. Girbea, C. Suciu, S. Nechifor, and F. Sisak, "Design and implementation of a service-oriented architecture for the optimization of industrial applications," *IEEE Trans. Ind. Informat.*, vol. 10, no. 1, pp. 185–196, Feb 2014.



Václav Jirkovský (M'14) received the M.Sc. (2010) from Czech Technical University in Prague, Czech Republic. He is with Czech Institute of Informatics, Robotics, and Cybernetics, Czech Technical University as Researcher and Rockwell Automation R&D Center as Research Engineer. Currently, he is working towards the Ph.D. degree at the Czech Technical University. His interests include ontologies, Big Data, and AI. He has co-authored around 16 publications and 2 granted patents.

Marek Obitko (M'07) received his Ph.D. (2007) in Artificial Intelligence and Biocybernetics from Czech Technical University in Prague, Czech Republic. He is with Rockwell Automation R&D Center in Prague as Senior Research Engineer and R&D Team Leader. His interests include knowledge representation, ontologies, semantic web, collaborative engineering and security. He has co-authored around 30 publications and 5 granted patents. He serves as a referee for conferences and journals.



Vladimír Mařík (M'95) received his PhD. in 1979 and DrSc. degree in 1989 from the Czech Technical University in Prague. He acts as the Director of the Czech Institute of Informatics, Robotics, and Cybernetics (CIIRC) at the Czech Technical University. His main professional interests include distributed AI, multi-agent systems, planning and scheduling for manufacturing, etc. He is author or coauthor of 7 monographs, 8 textbooks, more than 110 papers at international conferences. He acts as the VP of the IEEE SMC since 2014.