Integration and massive storage of hydro-meteorological data combining big data & semantic web technologies

Andrés Tello, Renán Freire, Mauricio Espinoza, Víctor Saquicela

Computer Science Department, University of Cuenca, Av. 12 de abril, Cuenca, Ecuador .

Autores para correspondencia: andres.tello@ucuenca.edu.ec, renan.freire@ucuenca.edu.ec, mauricio.espinoza@ucuenca.edu.ec, victor.saquicela@ucuenca.edu.ec

Fecha de recepción: 28 de septiembre 2015 - Fecha de aceptación: 12 de octubre 2015

ABSTRACT

Ecuador contains an immense collection of hydro-meteorological data, informing us via standards how to locate and invoke them. If we want to make such data easier to understand and use, we need to store them in a common repository and annotate them by means of descriptive metadata. This paper proposes an approach for the massive storage, integration and semantic annotation of hydro-meteorological data using an open source integration container, NoSQL databases and Semantic Web technologies. The main contributions of this paper are: i) a shared common repository of hydro-meteorological data, ii) automatic semantic annotation to formally describe the data sources, and iii) efficient mechanisms for searching and retrieving hydro meteorological data.

Keywords: Hydro-meteorological data, data integration, big data, semantic web, NoSQL.

RESUMEN

Ecuador contiene una inmensa colección de datos hidro-meteorológicos usualmente descritos usando estándares que nos indican cómo localizarlos y cómo invocarlos. Si queremos hacer que esos datos sean potencialmente más sencillos de entender y usar se requiere almacenarlos en un repositorio común y anotarlos formalmente usando metadatos descriptivos. Este artículo propone un mecanismo para el almacenamiento masivo, integración y anotación semántica de datos hidro-meteorológicos utilizando un framework de integración, bases de datos NoSQL y tecnologías de web semántica. Las contribuciones principales de este artículo son: i) Un repositorio compartido de datos hidro-meteorológicos, ii) anotación semántica automática para describir las fuentes de manera formal, y iii) mecanismos eficientes de búsqueda y consulta de datos hidro-meteorológicos

Palabras clave: Datos hidro-meteorológicos, integración de datos, big data, web semántica, NoSQL.

1. INTRODUCTION

Today, Ecuador is generating a vast amount of hydro-meteorological data. Institutions such as INAMHI[1], SENAGUA[2], PROMAS[3], among others, manage meteorological monitoring stations located in different regions of Ecuador. Those institutions and field-related research groups own repositories of meteorological data whose size is increasing dramatically hindering its proper management, usage and exploitation. Some of these repositories are managed almost manually, i.e., data are stored in digital media which do not facilitate its usage. In addition, the lack of proper data management tools, and the need for searching and querying tools do not facilitate researchers to

[1] INAMHI (Instituto Nacional de Meteorología e Hidrología): http://www.serviciometeorologico.gob.ec/

[2] SENAGUA (Secretaria Nacional del Agua): http://www.agua.gob.ec/

[3] PROMAS (Programa para el Manejo del Agua y del Suelo): http://promas.ucuenca.edu.ec

navigate through the repositories and to use the data for their own experiments (e.g., simulations, predictions, real time analysis, etc.).

When people working with hydro-meteorological data are also the owners of the data sources, integration is easier to achieve. They may define, beforehand, a unique data model to represent the generated data. Thus, every monitoring station or any other data source would generate the data in the same format. However, hydro-meteorological data and particularly data from sensors are highly heterogeneous, and coping with such heterogeneity is challenging and time consuming (Aberer *et al*., 2006). Each monitoring station uses its own data formats (e.g., free text, spreadsheets, relational databases, reports, etc.). In addition, each deployment has its own system for gathering, processing, and publishing the data coming from sensors, especially if they are managed by different authorities (e.g., INAMHI, ETAPA, PROMAS, etc.).

The current situation when working with hydro-meteorological data is that each repository is managed as a silo (see Fig. 1). Researchers working with meteorological data do not know the data available outside their enterprise or organization. Researchers wonder whether repositories are available, how to find them, how to access the data, and whether they have to pay for the data, etc. In addition, even within an organization not all the people know what kind of data is available for experimentation. Their main concerns are about the data formats, available tools, querying mechanisms, etc.

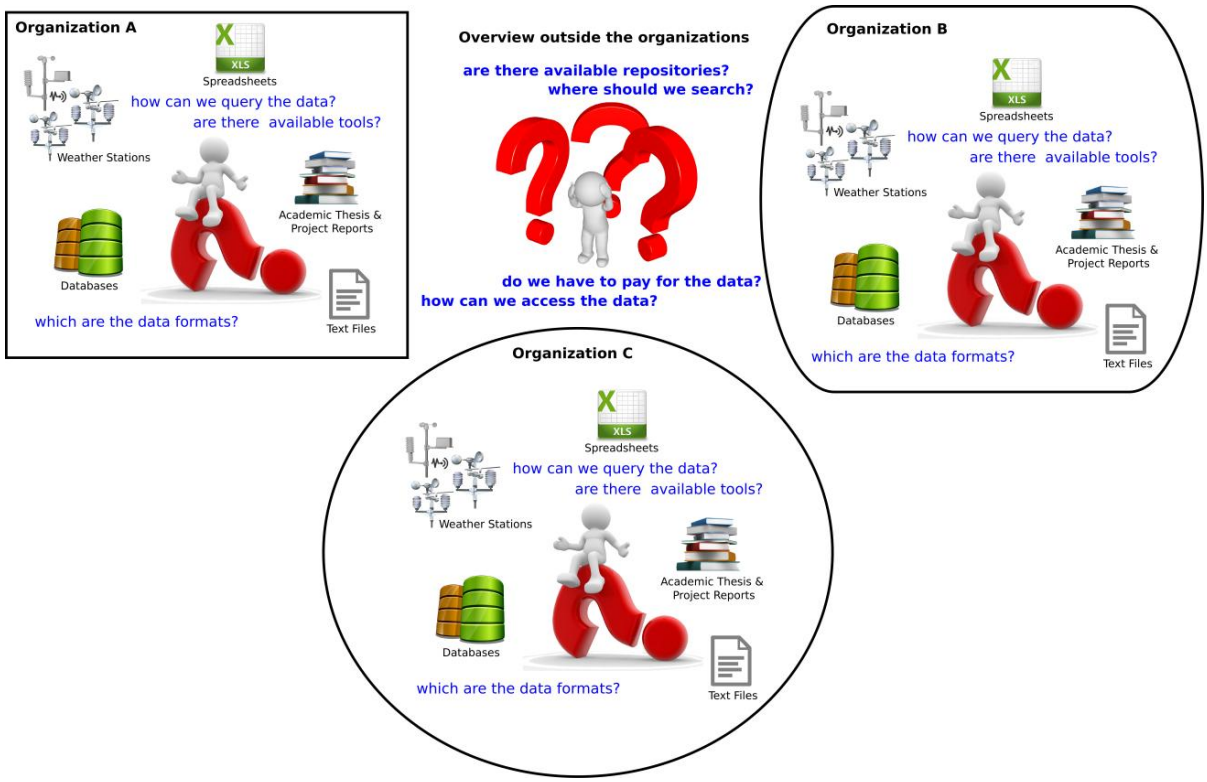

Figure 1. Typical situation when working with hydro-meteorological data.

From the aforementioned situation we identified three main concerns: 1) the need of a common shared hydro-meteorological data repository which can be used by researchers from academia and industry, 2) the need of an automatic semantic annotation mechanism which allows to formally describe the data sources, and 3) the need to provide to the end-users the tools for accessing and exploiting the data. We propose a data integration platform for collecting and storing the hydro-meteorological data combining the theory and the technologies from Big Data (Villars *et al*., 2011) and Semantic Web **(**Berners-Lee *et al*., 2001). This fusion allows alleviating the heterogeneity and associated complexity, thereby tackling simultaneously the previously identified problems.

Working with hydro-meteorological data means dealing with huge data volumes. Datasets easily reach terabytes[4] of data in few days or even hours. In addition, it involves dealing with quite diverse datasets. Each hydro-meteorological monitoring station generates the data using its own data formats. Moreover, the data are generated at a high frequency. The sensors at the hydro-meteorological stations may generate new data every second. As described, hydro-meteorological data possesses three main features: volume, variety, and velocity; this is so called "the 3 Vs" of a Big Data repository (Sagiroglu & Sinanc, 2013). Hence, Big Data technologies involve different concepts, tools, and mechanisms that allow facing the management-complexity associated to the datasets.

On the other hand, the use of Semantic Web principles and technologies facilitates the creation of conceptual models to formally describe the datasets by means of ontologies[5]. In addition, Semantic Web facilitates sharing and reusing data (Atemezing *et al*., 2013). Using ontological models has as advantage that researchers and developers, using e.g. hydro-meteorological data, do not have to understand the data format, or work on data transformations because the data would be already standardized. Furthermore, using ontologies for representing the data, the datasets might be extended or combined with datasets from other domains. For instance, hydro-meteorological datasets might be combined with agricultural and/or touristic datasets to find for example correlations between them or to make impact studies that benefit these sectors.

A hydro-meteorological data integration platform is the starting point for projects that base its results on this type of data. Without a platform, that allows researchers free access and exploitation of the hydro-meteorological data available in Ecuador, we are far from transforming such data in new knowledge - the foremost goal of scientific research.

The remainder of this paper is divided as follows: Section 2 describes similar works made by other researchers; Section 3 presents the authors proposal for a hydro-meteorological data repository and present some results of experiments with focus on the assessment of repositories; and Section 4 presents the conclusions and provide some ideas of future work.

2. RELATED WORK

There is not that much literature available on the combined use of Big Data and Semantic Web to facilitate the integration, storage, and exploitation of hydro-meteorological data. However, quite some research has been published detailing the need to integrate these heterogeneous data sources in a platform that allows storing and exploiting such data more efficiently. In this section we present what other people have been doing in this area.

Atemezing *et al*. (2013) proposed the transformation, publishing and exploitation of the data provided by the Spanish Meteorology Public Agency (AgenciaEstatal de Meteorología de España). The meteorological data are in CSV format which are then transformed to RDF using the W3C ontology SSN[6]. Then, these data are published in the Linked Open Data[7] (LOD) cloud for its further exploitation. This work uses Semantic Web to annotate the data with respect to an ontology which facilitates and enhances its later exploitation. However the authors mention that one of the main limitations is the lack of mechanisms to tackle the accelerated increase in size of the datasets. The data sample for this study was generated every ten minutes from 250 monitoring stations. This frequency of data generation produces 10'488'672 records which have to be processed daily. Such fast size growing of the datasets forces the authors to keep only the data generated in the last week. Our proposal of

[4] The **terabyte (TB)** is a multiple of the unit byte for digital information. 1 TB = 10^3 GB = 10^6 MB = 10^9 KB = 10^{12} bytes

[5] In computer science and information science, an ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse.

[6] SSN ontology. http://www.w3.org/2005/Incubator/ssn/ssnx/ssn

[7] Linked Data. http://linkeddata.org/

combining Big Data and Semantic Web may avoid the limitations caused by the fast growing size of the datasets, and take advantage of the key features from both areas.

Patni *et al.* (2010) present a framework to transform and publish the data from the Meteorology Department of the University of Utah and make them publicly available in the LOD cloud. The data are transformed to RDF following the OGC[8] standards and the resulting datasets are published in the cloud. Patni *et al*. (2011) propose an extension to previous work extracting the features from the RDF data. Such features allow them to detect and describe natural phenomena (e.g., storms, hurricanes, etc.). These are other examples of the use of Semantic Web to enrich and leverage the hydro-meteorological data coming from sensors. However, these authors did not mention how they deal with scalability, neither how they manage the unavoidable increase of datasets volume.

People working with hydro-meteorological data coming from sensors realized that the volume of data volume to be stored and processed is increasing dramatically. As stated before, this kind of datasets can be considered as Big Data. Bifet (2013) says that instead of associating Big Data only with huge datasets of a specific size (e.g., petabytes), we should think of datasets which cannot be managed without the use of new technologies and new algorithms. This motivated researchers working on Semantic Web to combine their area of expertise with Big Data.

Cudre-Mauroux *et al.* (2013) made a study of Big Data tools to process semantic data, i.e., data in RDF format. They describe four types of NoSQL databases and execute different performance tests which are compared against the results obtained from using traditional RDF triple stores. The authors concluded that distributed NoSQL databases may produce similar results and sometimes overcome the problems when using a native RDF triple store (e.g., 4Store). For instance, in a trial using 1 billion of triples and 16 nodes, they obtained better results using the NoSQL databases Cassandra and Jena+HBase than those using 4Store. In addition, they observed that the loading time of triples to the server scales more naturally when using NoSQL databases running in parallel.

Other studies are focused on the creation of tools that allow managing huge volumes of semantic data using Big Data technologies. Cuesta *et al*. (2013) proposed architecture to process semantic data combining batch and real time processing. In their approach they separated the management of large data volumes from the generation and exploitation of the data in real time. The data are compressed using RDF/HDT[9] and stored, while the real time processing requirements are handled using NoSQL databases.

Zeng *et al*. (2013) presented Trinity.RDF, a distributed memory-based graph engine for large volumes of RDF data. They proposed to store RDF data in its native form, i.e., graphs. This approach shows a better performance for SPARQL queries with respect to traditional RDF triple stores (sometimes orders of magnitude better). In addition, they stated that since the data are stored as a graph, the systems can support other kind of operations, e.g., random walks, reachability, etc.

Different approaches presented in this section focus, from one hand, on the use of Semantic Web to model hydro-meteorological data, and from the other hand a different point of view on the need of using Big Data to manage RDF data. Nonetheless, all the authors agree that one of the main challenges today is how to deal with the steep growing of the size of datasets. Our proposal is focused on combining both areas, Big Data and Semantic Web, to provide an integral solution for the integration, storage and exploitation of hydro-meteorological data.

[8] The Open Geospatial Consortium (OGC) is an international industry consortium of 519 companies, government agencies and universities participating in a consensus process to develop publicly available interface standards.

[9] HDT (Header, Dictionary, Triples) is a compact data structure and binary serialization format for RDF that keeps big datasets compressed to save space while maintaining search and browse operations without prior decompression.

3. SHARED HYDRO-METEOROLOGICAL DATA REPOSITORY

In this section, the authors present their platform for the integration and massive storage of hydro-meteorological data into a shared repository. Our proposal combines the tools and principles from Big Data and Semantic Web to overcome the problems described in Section 1. The proposed platform relies on an open source integration container (Apache ServiceMix[10]) and different Big Data and Semantic Web technologies (e.g., Apache Hadoop[11], Apache Marmotta[12], Apache Cassandra[13], etc.). Figure 2 shows an overview of the platform.

The top layer (see (A) in Fig. 2) represents the hydro-meteorological data sources. In our proposal we assume that the data are stored electronically and that they are accessible through a communication network, i.e., we do not deal with the data capturing but we consume the hydro-meteorological data available in any digital media. The data at the source side may be accessed and stored in different digital media, either in a database system, spreadsheet and text files within a file system, or accessible through a web service.

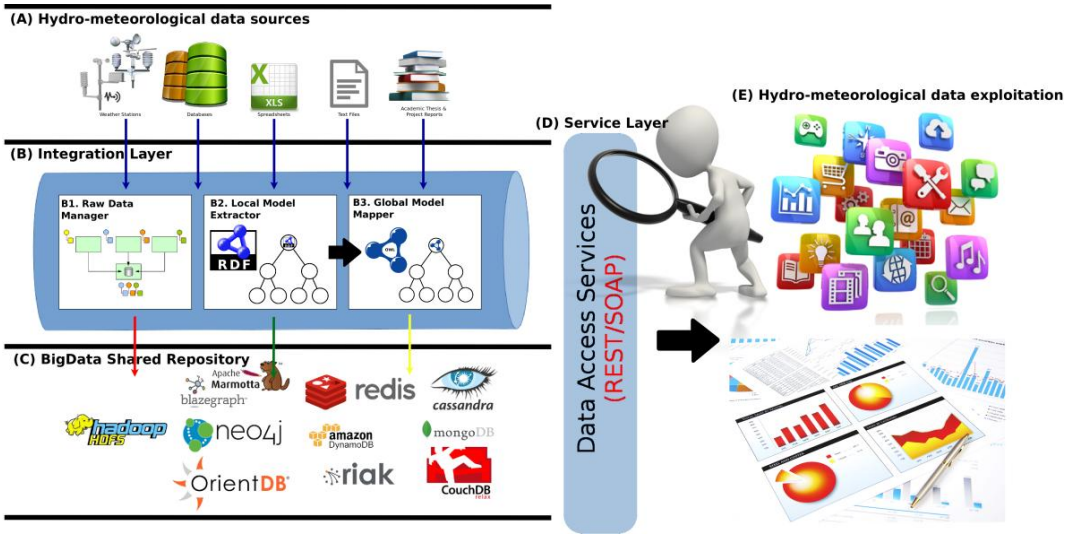

Figure 2. Hydro-meteorological data repository architecture.

The Integration Layer (see (B) in Fig. 2) coordinates the integration process. It serves as a mediator between the data sources and the shared data repository. Different processes or handled by this layer, e.g., data transformations, routing, messaging, etc. This layer is implemented on top of Apache ServiceMix, an integration container that includes several Apache technologies such as ActiveMQ[14]for messaging, Camel[15] for routing, the CXF[16] web services framework, and the OSGi[17] container Karaf[18] onto which various components and applications can be deployed. This layer contains three main components: (B1) the Raw Data Manager (RDM), (B2) the Local Model Extractor (LME), and (B3) the Global Model Mapper (GMM). The RDM component is responsible for gathering an exact copy of the raw data and after a pre-processing storing such data into the common repository. The LME component is responsible for extracting the main characteristics of the input data

[10] Apache ServiceMix. http://servicemix.apache.org/
[11] Apache Hadoop. https://hadoop.apache.org/
[12] Apache Marmotta. http://marmotta.apache.org/
[13] Apache Cassandra. http://cassandra.apache.org/
[14] Apache ActiveMQ: http://activemq.apache.org/
[15] Apache Camel: http://camel.apache.org/
[16] Apache CXF: http://cxf.apache.org/
[17] OSGi. http://www.osgi.org/Main/HomePage
[18] Apache Karaf: http://karaf.apache.org/

sources, e.g., what type of data are produced as output, extra metadata, how to invoke it, etc. The GMM component is responsible for transforming the local model to the global model, in this work, the global model is represented by an ontology in the hydro-meteorological domain (e.g., SSN[19], Meteorological sensor ontology[20], ISO 19156 Observation Model[21], etc.). The relationship between local and global model is performed through mappings, this allows the automatization of the transformation process.

The bottom layer (see (C) in Fig. 2) depicts the Big Data shared repository which stores the hydro-meteorological integrated data. Currently, we are using the Hadoop Distributed File System (HDFS), a distributed file system designed to run on commodity hardware (Borthakur, 2007), for storing the raw data. HDFS was designed to store very large datasets reliably (Shvachko *et al*., 2010). Since our repository is intended to store the hydro-meteorological data produced and located in different regions of Ecuador, we consider that this scenario is a suitable to use the HDFS. In addition, using HDFS, all the original raw data will be in one (logical) place allowing researchers to use such data in the way is more convenient for them. Let us illustrate this situation with an example. Hydro-meteorological data include time series and it is known that missing data points in time series data are very common (Fung, 2006). Therefore it may be necessary to estimate those missing data points using different methods such as for example interpolating with a Cubic Spline, autoregressive moving average models, Kalman filter, among others (Fung, 2006). Since the researchers will have access to the original data stored in a common repository they may apply different algorithms, including the algorithm which is most suited for their data. Moreover, they may tweak some algorithms are re-run the estimation process until the results are adequate to their needs.

Two different experiments were conducted to test the concept. In the first one, we consumed that the data are stored in a relational database and after a pre-processing, inside the Raw Data Manager within the integration platform, the data are stored into the HDFS. For this experiment we use Apache Sqoop[22], a tool for transferring bulk data between Apache Hadoop and structured data stores such as relational databases. In the second experiment we used files from a file system (e.g., spreadsheets, text, csv, etc.), which consequently were pre-processed within the integration platform, and finally the data were stored in the HDFS (see Fig. 3).

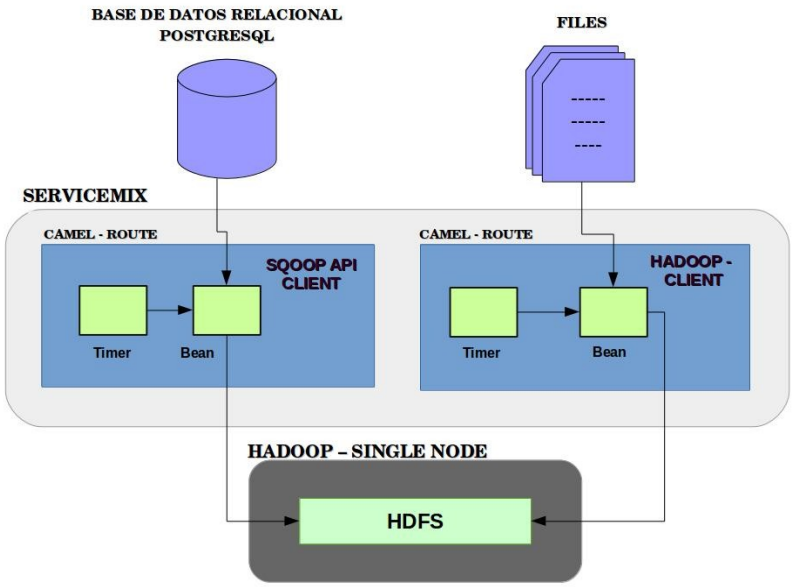

Figure 3. Tests executed as a proof of concept.

[19] Semantic Sensor Network Ontology. http://lov.okfn.org/dataset/lov/vocabs/ssn
[20] Meteorological sensor ontology. http://www.w3.org/2005/Incubator/ssn/ssnx/meteo/aws
[21] ISO 19156 Observation Model (om). http://lov.okfn.org/dataset/lov/vocabs/om
[22] Apache Sqoop. http://sqoop.apache.org/

Additionally, we compared the performance of different NoSQL databases which allow storing RDF and querying the data using standard SPARQL. Currently, we are testing Apache Marmotta using the BlazeGraph[23] Big Data back-end. BlazeGraph is a graph database which supports the Blueprints[24] and RDF/SPARQL APIs. Since RDF data linking structure forms yield labeled graphs (RDF Working Group 2014), a graph database is the most natural form for storing RDF data. Another alternative is the Neo4j graph database. Since Blueprints provides an implementation for both, BlazeGraph and Neo4j, we are testing both technologies.

Currently, we are working on the Data Access Services Layer (see (D) in Fig. 2) which is responsible for providing a querying interface for researchers and people interested on working with hydro-meteorological data in their experiments. Such web-based interface will allow creating different applications and tools (see (E) in Fig. 2) for the efficient exploitation of the hydro-meteorological data produced in different regions of Ecuador.

4. CONCLUSIONS & FUTURE WORK

A platform for the integration and storage of vast volumes of hydro-meteorological data into a shared common repository is presented. A combination of two technologies, Big Data and Semantic Web, is proposed. On one hand, Big Data technologies help to face the challenges of dealing with velocity, variety, and volume, some intrinsic characteristics of hydro-meteorological data. On the other hand, Semantic Web allows facing the heterogeneity problem described formerly by means of ontologies. Ontologies provide a vocabulary which defines the concepts and relationships used to describe and represent the hydro-meteorological data. In addition, a test was conducted using two different scenarios. In both of them we conclude that it is feasible to integrate heterogeneous data sources of hydro-meteorological data into a shared common repository using state of the art technologies (e.g., Apache Service Mix, HDFS, etc.) which complies with Big Data principles.

One of the main limitations of our approach is the lack of a formal evaluation of the proposed integration platform. In future, we will perform a more extensive evaluation as to assess the feasibility of our approach. Such evaluation includes the execution of several stress tests to assess the performance of the platform when working with heavy loads of hydro-meteorological data. Parallel herewith, we plan to evaluate other NoSQL technologies which might be more suitable for managing hydro-meteorological data (e.g., Cassandra, HBase, etc.). Such tests will enable to better choose the right technology as the Big Data Back-End in the final version of the proposed integration platform.

REFERENCES

Aberer, K., M. Hauswirth, A. Salehi, 2006. *The Global Sensor Networks middleware for efficient and flexible deployment and interconnection of sensor networks*. ACM/IFIP/USENIX 7th International Middleware Conference, EPFL, Lausanne, Switzerland.

Corcho, Ó., D. Garijo Verdejo, J. Mora, M. Poveda Villalon, D. Vila Suero, B. Villazón-Terrazas, G.A. Atemezing, 2012. *Transforming meteorological data into linked data*. Undefined, 1, 1-5, IOS Press. Available at http://www.semantic-web-journal.net/sites/default/files/swj281_0.pdf.

Berners-Lee, T., J. Hendler, O. Lassila, 2001. The semantic web. *Scientific American*, 284(5), 28-37.

Bifet, A., 2013. Mining big data in real time. *Informatica*, 37(1).

Borthakur, D., 2007. *The hadoop distributed file system: Architecture and design*. Hadoop Project Website, 11, 21.

[23] BlazeGraph. https://www.blazegraph.com/product/
[24] Blueprints. https://github.com/tinkerpop/blueprints/wiki

Cudré-Mauroux, P., I. Enchev, S. Fundatureanu, P. Groth, A. Haque, A. Harth, F.L. Keppmann, D. Miranker, J.F. Sequeda, M. Wylot, 2013. *Nosql databases for rdf: an empirical evaluation*. In: The Semantic Web-ISWC conference, 310-325.

Cuesta, C.E., M.A. Martínez-Prieto, J.D. Fernández, 2013. Towards an architecture for managing big semantic data in real-time. *Software Architecture*, 45-53.

Fung, D.S.C., 2006. Methods for the estimation of missing values in time series. PhD thesis, Edith Cowan University, Perth, Australia.

Patni, H., C. Henson, A. Sheth, 2010. *Linked sensor data*. In: Collaborative Technologies and Systems (CTS). IEEE International Symposium, 362-370.

Patni, H., C.A. Henson, M. Cooney, A.P. Sheth, K. Thirunarayan, 2011. Demonstration: real-time semantic analysis of sensor streams. Available at http://corescholar.libraries.wright.edu/cgi/viewcontent.cgi?article=1249&context=knoesis.

RDF Working Group, 2014. Resource Description Framework (RDF). Available at https://www.w3.org/2001/sw/wiki/RDF.

Sagiroglu, S., D. Sinanc, 2013. *Big data: A review*. In: Collaboration Technologies and Systems (CTS), IEEE International Conference, 42-47.

Shvachko, K., H. Kuang, S. Radia, R. Chansler, 2010. *The hadoop distributed file system*. In: Mass Storage Systems and Technologies (MSST), IEEE 26th Symposium, 1-10.

Villars, R.L., C.W. Olofson, M. Eastwood, 2011. *Big data: What it is and why you should care*. White Paper, IDC. Available at http://www.admin-magazine.com/HPC/Vendors/AMD/Whitepaper-Big-Data-What-It-Is-and-Why-You-Should-Care.

Zeng, K., J. Yang, H. Wang, B. Shao, Z. Wang, 2013. *A distributed graph engine for web scale rdf data*. Proceedings of the VLDB Endowment, 6(4), 265-276.