

From the sensor data streams to linked streaming data. A survey of main approaches

K. R. Llanes¹, M. A. Casanova¹, N. M. Lemus²

¹ Pontificia Universidade Católica do Rio de Janeiro, Brazil

kllanes@inf.puc-rio.br, casanova@inf.puc-rio.br

² Laboratório Nacional de Computação Científica, Brazil

noelml@lncc.br

Abstract. Nowadays, large amounts of data are produced by sensor networks. They are continuously producing information about real world phenomena in the form of data streams. However, these data are generated in raw and different formats, lacking the semantics to describe their meanings, which imposes barriers in accessing and using them. To tackle this problem several solutions using Linked Data Principles have been proposed. In this paper, we survey the main solutions developed by the research communities for publishing stream data in the Web of Data, identifying their strengths and limitations. Over that basis, the main steps that someone should follow to publish data streams in a manner that anyone can use them, with a minimal understanding the details, are defined; which represents the main contribution of this work. We also highlight the main challenges that emerge from this survey, concluding with a list of research tasks for future work.

Categories and Subject Descriptors: H.2.5 [**Database Management**]: Heterogeneous Databases; C.2.3 [**Computer-Communication Networks**]: Network Operations

Keywords: data streams, linked data, semantic web, sensor data publishing

1. INTRODUCTION

In recent years, data sensor networks have been deployed in various domains (medical sciences for patient care using biometric sensors, wildfire detection, meteorology for weather forecasting, satellite imagery for earth and space observation, agricultural lands, etc.). The sensors are distributed across the globe, capturing and continuously producing an enormous amount of data about a number of real world phenomena in the form of data streams.

However, typically, the data produced by sensor networks are in raw and different formats, lacking the semantics to describe their meaning. This failure intensifies the current traditional problem "too much data and not enough knowledge" [Sheth et al. 2008] and imposes barriers in accessing and using sensor data in applications and linking them with other related data sources.

To tackle this problem, several solutions using Linked Data Principles [Berners-lee 2006] have been proposed. They allow integrate sensor technologies with Semantic Web technologies in order to publish sensor data streams in an enriched and standardized way, so that they can be accessed and consumed by external applications. The publication process consists of transforming the data streams into linked streaming data following the Linked Data Principles.

During the process of publishing data streams in the Linked Open Data (LOD) cloud, the time component plays a key role, which substantially changes the way of data processing compared with the publishing of static data.

Copyright©2014 Permission to copy without fee all or part of the material printed in JIDM is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

The publishing of static data in the LOD cloud is composed of several activities: specification, modelling, generation, publication and exploitation. The specification refers to a preliminary set of tasks to identify and analyse the data to be published. Then, we need to select the ontology or ontologies to be used for modelling and semantically describing data. After that, the data are transformed to standard representation in RDF format [<http://www.w3.org/RDF>] and linked to external data sources through generation activity. This activity ends with a meaningful and enriched tripleset. During the publication activity, this enriched tripleset is stored and published in a triple store to be consumed later. Once data are published on the Web of Data, they would be queried and consumed. The activity that takes care of these tasks is called exploitation, which is the main purpose of the publication process as whole.

In multiple domains, the time component is critical to make the right decisions quickly. In terms of data stream processing, it implies that it is done in real-time. That means data from sensor observations should be processed on-the-fly with a minimum delay. To fulfill this requirement, significant modifications to the traditional static data publishing process should be made, such as the incorporation of data compression, data stream abstraction, continuous queries and the generation of links in real time, among others.

Several efforts have been developed for publishing data streams on the Web of Data. Some take into account the real-time and others do not consider it. In this paper, we survey the main works available in the research literature, identifying their weaknesses and strengths. Based on an analysis of the strengths and weaknesses, we propose a set of next research tasks to be facet in the near future.

The remainder of the paper is organized as follows. Section 2 describes in detail the main steps that someone should follow to publish data streams in a way that anyone can use them with minimal understanding of the data details. Section 3 shows the most relevant approaches proposed to publish data streams on the Semantic Web following the Linked Data Principles. Section 4 discusses lessons learned and open challenges that emerged from this survey. Section 5 concludes the paper and presents our future research directions.

2. SENSOR DATA PUBLISHING ON THE SEMANTIC WEB

Sensor networks employ various types of hardware and software components to observe and measure physical phenomena and make the obtained data available through different networking services. Applications and users are typically interested in querying various events and requesting measurement and observation data from the physical world. Through the process of publishing sensor data on the Semantic Web, knowledge is added over raw sensor data in order to satisfy the high-level information requested by the queries.

The process of publishing sensor data in the Semantic Web encompasses three main stages: mapping and conversion from data streams to RDF streams, storing RDF streams and linking them with related data sources existing in the LOD cloud. To carry out this process a set of important tasks are required, such as: *(i) selection of ontologies to semantically describe data streams; (ii) defining the mapping language to do the conversion; (iii) selection of continuous query languages; (iv) choosing the appropriated datasets from the LOD cloud to create the links.* See Figure 1. To support the complete process a stream publishing framework is being developed.¹

2.1 Selection of ontologies

With the development of semantic sensor networks a number of ontologies describing the sensor network domain have been brought forth in the past years. A detailed survey was performed by

¹<https://github.com/nmlemus/streams2LSD>

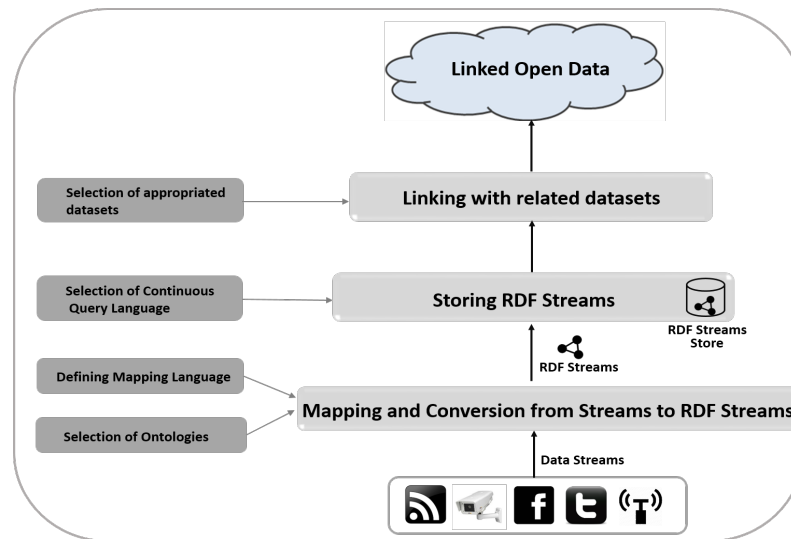


Fig. 1. Data Stream Publication Process

Michael Compton et al. in [Compton et al. 2009], where eleven sensor network ontologies were analyzed. Therefore, considering the need of standardization regarding sensor networks ontologies, the Semantic Sensor Network Incubator Group from W3C was formed, with the purpose of developing ontologies for sensor networks and searching for appropriate methods for enhancing available standards with semantic technologies. As a result of the efforts of this group, the Semantic Sensor Network (SSN) ontology [Compton et al. 2012] was defined, which can describe capabilities, measurements and resultant observations from sensors.

The W3C Semantic Sensor Network Incubator Group also developed a methodology to perform semantic annotations over the data generated by sensors following the standards defined by the Open Geospatial Consortium (OGC). These standards help describe observed phenomena such as space, time and theme.

Spatial metadata provide information regarding the sensor location and data, in terms of either a geographic reference system, local reference, or named location. Temporal metadata provide information regarding the time instant or interval when the sensor data is captured. Thematic metadata describe a real world state from sensor observations, such as objects or events. All these metadata play an essential role in managing sensor data and provide more meaningful descriptions and enhanced access to sensor data. Both projects developed by W3C Semantic Sensor Network Incubator Group: the SSN ontology (SSNO) and the proposed methodology, facilitate the stream data semantic fusion applications and the integration of stream data with linked data sets, because the fact does not only publish the streaming data, but also integrate them with other related datasets. Sometimes, sensor ontologies are not able to provide all the semantics needed by a scientific system and additional ontologies are often required.

2.2 Defining the mapping language

Several languages have been proposed by the Semantic Web research communities for expressing customized mappings from relational databases to RDF datasets. Such mappings provide the ability to view existing relational data in an RDF data model, expressed in a structure and target vocabulary of the mapping author's choice. D2R [Bizer 2003], R2O [Barrasa et al. 2004] and R2RML [Consortium 2012] that represents the standard language are some of them. They are fruitfully on transforming

static relational data to RDF, but present some disadvantages to face the challenge of converting data streams to RDF streams.

Despite the existence of this gap, solutions for streaming data mapping and querying using ontology-based approaches have been little explored. Calbimonte et al. [Calbimonte et al. 2010] presented an extension of R2O called S2O for the data stream to RDF mapping. Also, Harth et al. [Harth et al. 2013] developed an extension for R2RML with the same purpose. These last extensions are better suited to support the publishing of stream data.

2.3 Selection of continuous queries languages

Languages such as SPARQL are designed to execute queries over RDF triples, but they do not have the functionalities to query RDF streams. To face this challenge, continuous RDF query languages have been proposed.

Barbieri et al. [Barbieri et al. 2009] introduced Continuous SPARQL (C-SPARQL) as the extension of SPARQL for querying RDF streams. It supports continuous queries, registered and continuously executed over RDF data streams, considering windows of such streams. C-SPARQL is currently not designed to handle large volumes of data, which constitutes their main weakness.

SPARQLstream [Calbimonte et al. 2011] is an extension to SPARQL for RDF streams. It has been inspired by previous proposals C-SPARQL and SNEEQL [Brenninkmeijer and Galpin 2008], but with some improvements: it only supports windows defined in time; the result of a window operation is a window of triples, not a stream, over which traditional operators can be applied. It uses S2O and R2RML for the definition of stream-to-ontology mappings. Its main disadvantage is that currently does not support querying on both stream and RDF datasets.

Ainic et al. [Anicic and Fodor 2011] developed Event Processing SPARQL (EP-SPARQL). It is a continuous query language that uses a black box approach backed by a logic engine. It translates queries into logic programs which are then executed by a Prolog engine. EP-SPARQL provides a unified execution mechanism for event processing and stream reasoning which is grounded in logic programming. The main deficiency of EP-SPARQL is that its performance drops significantly for complex queries.

Le Phuoc [Phuoc 2013] presented Continuous Query Execution over Linked Stream (CQELS), an adaptive execution framework for Linked Stream Data and Linked Data. CQELS provides a flexible architecture for implementing efficient continuous query processing engines over Linked Data Stream and Linked Data.

To the best of our knowledge, the most complete approach to continuous queries over RDF streams is CQELS. Despite its scalability issues with respect to multiple concurrent queries, the CQELS engine can achieve better performance than other black box systems in order of magnitude. It represents a solution for RDF stream processing built on top of the notion of linked stream data. The solution offers a native way to interpret and implement common stream processing futures (time window operator, relational database like join and union operators, and stream generation operator) in an RDF data stream processing environment.

2.4 Choosing LOD datasets to create the links

Another important task in the process of data stream publishing on the Semantic Web is the selection of the most suitable triplesets with which RDF streams may be interlinked. It will allow users to take advantage of existing knowledge. Once the most suitable triplesets are found, the next step is to link them with a local sensor tripleset, thus completing the process of publishing. However, interlinking is a laborious task. Thus, users interlink their triplesets mostly with data hubs, such as DBpedia

and Freebase, ignoring the most specific yet often even more promising triplesets. To alleviate this problem, some triplement recommendation tools have been implemented.

Lopes et al. [Lopes et al. 2013] presented a triplement recommendation's approach using strategies borrowed from social networks. To generate the ranked list, the procedure uses a recommendation function adapted from link prediction measures used in social networks. The tool obtains high levels of recall and reduces in up to 90% the number of triplesets to be further inspected for establishing appropriate links.

Caraballo et al. [Caraballo et al. 2014] presented a Web-based application, called TRTML, that explores metadata available in Linked Data catalogs to provide data publishers with recommendations of related triplesets. TRTML combines supervised learning algorithms and link prediction measures to provide recommendations. Its high precision and recall results demonstrate the usefulness of TRTML.

Lopes et al. [Lopes et al. 2014] developed RecLAK. It is a Web application developed for the LAK Challenge 2014 focused on the analysis of the LAK dataset metadata and provides recommendations of potential candidate datasets to be interlinked with LAK dataset. RecLAK follows an approach to generate recommendations based on Bayesian classifiers and on Social Networks Analysis measures. Furthermore, RecLAK generates graph visualizations that explore the LAK dataset over other datasets in the LOD cloud.

The main disadvantage of these triplement recommendation tools is that they are not able to do the recommendation process on-the-fly, since they are not designed to act in real-time, which represents a gap in the process of sensor's data publishing. For this reason, the current solution is to choose the LOD datasets related with each new sensor that will be incorporated into the sensor network, using the tools described above, before sensors start to capture observations.

3. PLATFORMS FOR SENSOR DATA PUBLISHING ON THE SEMANTIC WEB

Although the main goal of Linked Stream Data is to make available the sensor data in the LOD cloud, in real-time, quite a few projects have achieved it. In this section the most recently efforts of research communities to publish sensor data on the Semantic Web will be analyzed. Some of them do not publish sensor data in real-time, which is its main weakness, but may serve as starting point for future work.

3.1 Non real-time approaches

Le-Phuoc et al. [Phuoc and Hauswirth 2009] presented an approach and an infrastructure which makes sensor data available following the Linked Open Data principles and enables the seamless integration of such data into mashups. This project publishes sensor data as Web data sources which can then be easily integrated with other Linked Data sources and sensor data. Also, it allows users to describe and annotate semantically raw sensor readings and sensors. These descriptions can then be exploited in mashups and in Linked Open Data scenarios and enable the discovery and integration of sensors and sensor data at large scale. The user-generated mashups of sensor data and Linked Open Data can in turn be published as Linked Open Data sources and be used by other users.

Patni et al. [Patni et al. 2010] presented a framework to make this sensor data openly accessible by publishing it on the LOD cloud. This is accomplished by converting raw sensor observations to a standard representation in Resource Description Framework (RDF) and linking with other datasets on LOD. With such a framework, organizations can make large amounts of sensor data openly accessible, thus allowing greater opportunity for utilization and analysis. They were the first to add to the LOD cloud a large dataset of sensor descriptions and measurement, by first representing it in Observation and Measurements (O&M) standard.

Barnaghi and Presser [Barnaghi et al. 2010] proposed a platform called Sense2Web for publishing sensor data description defined by spatial, temporal and thematic attributes. The platform offers an interface for publishing linked sensor data without requiring from the users a semantic technological background. The sensor observation and measurement data can also be published following similar principals. However, the publishing of observation and measurement data raises other concerns such as time-dependency, scalability, freshness and latency.

Moraru et al. [Moraru et al. 2011] proposed a system for publishing sensor data following the Linked Data Principles and providing hereby integration with the Semantic Web. In their proposal they focused on a single sensor source and for storing sensor data they used a relational database, which represents its main deficiency; because a relational database is not prepared to support the continuous arriving of data streams.

3.2 Real-time approaches

Barbieri et al. [Barbieri and Valle 2010] proposed an approach to publish data as Linked Data streams. The approach uses C-SPARQL to register and run continuous queries over streams of RDF and C-SPARQL engine to publish the retrieved data as Linked data streams. To represent RDF in RDF streams, they proposed the use of two named graphs: stream graph (S-graph) and instantaneous graph (I-graph). An RDF stream can be represented using one s-graph and several i-graphs, one for each timestamp. The main limitation of this approach is that is only a prototype, and it does not have a finished application that supports it.

Le Phuoc [Le-Phuoc et al. 2011] proposed a Linked Stream Middleware, a platform to facilitate publishing Linked Data Stream and making it available to other applications. It provides the following functionalities: wrappers to access stream data sources and transform the raw data into Linked Stream Data, data annotation and visualization through a web interface and life querying over unified Linked Stream Data and data from the LOD cloud. Besides the processing of real-time data, it is also necessary to store the data generated, either for queries defined over a time period or for archiving purposes. It is here where it appears the main limitations of this approach: the triple storage cannot efficiently handle high update rates; the materializing sensor reading into triples is also inefficient, especially numeric readings and also, it runs into performance issues with complex queries.

Hasemann et al. [Hasemann et al. 2012] proposed Platform-independent Wiselib RDF Provider for embedded Internet of Thing (IoT) devices such as sensor nodes. It enables the devices to act as semantic data providers. They can describe themselves, including their services, sensors, and capabilities, by means of RDF documents. The greatest contribution of this proposal is the introduction of Streaming HDT, a lightweight serialization format for RDF documents that allows for transmitting compressed documents with minimal effort for the encoding. Also a platform allows to publish and share sensor data with reduce level of cost, less complexity of sensor data integration, and easy to access the integrated sensor data.

Harth et al. [Harth et al. 2013] developed a Web architecture that enables (near) real-time access to data sources in a variety of formats and access modalities. Also, it enables rapid integration of new live sources by modeling them with respect to domain ontology and automatically transforming all the arrived data streams from their format (CSV, TSV, JSON) to RDF and publishing them following the Linked Data Principles. This approach is a very good approximation to solve the problem related to the integration of new sensor devices into the LOD cloud, but it is still immature project.

4. LESSON LEARNED AND OPEN CHALLENGES

4.1 Lesson Learned

The publishing of data streams on the LOD cloud should take into account the following observations:

- (1) **Ontology selection:** There are several ontologies designed to semantically describe sensor data that help us during the annotation process. However, sometimes sensor ontologies are not able to provide all the semantics and additional ontologies are often required.
- (2) **Data stream selection:** Before starting the transformation process from data streams to RDF, it is extremely important to make an abstraction of streams to select the most significant data and not spend time to process those less relevant streams.
- (3) **Data compression:** An efficient and lightweight serialization format for RDF should be used, in order to transmit compressed documents with minimal effort for encoding.
- (4) **Data Integration:** Integrate information from heterogeneous sources (sensor networks and social networks) in order to support decision making in real-time. Integrating sensor data with data from social networks, allows you to capture human perception, implying that better decisions are made.

4.2 Open Challenges

In order to integrate sensor technologies with Semantic Web technologies and publish them as Linked Data Streaming in real-time, some efforts have been made. Nevertheless, some challenges are still being faced:

- (1) To publish and consume data from sensor data streams in real-time it is necessary a lighter mapping language, capable of guaranteeing on demand mapping and efficient conversion from sensor data streams to RDF streams.
- (2) The conversion of the data streams to RDF streams must be on-the-fly. This restriction captures the idea that the data must be continuously triplified, albeit with limited delay. To fulfill this requirement, techniques for efficient triplification should be developed.
- (3) The interlinking process of RDF streams with data sources of the LOD cloud must be on-the-fly. To address the restriction of minimum delay, interconnection techniques should be based on a strategy of preprocessing or caching data to accelerate the creation of links.
- (4) The lack of an efficient RDF storage that supports the real-time stream processing. Although the classical RDF storages are efficient to store RDF that will persist over time, they are not efficient to handle the RDF streams, because the RDF streams need to be stored, accessed and processed on-the-fly.

5. CONCLUSIONS AND FUTURE WORK

Real-time publishing of sensor data streams based on semantic technologies is indeed not only possible, but useful in many application areas. In this paper, we presented a study that covers the main approaches proposed to publish the sensor data in the LOD cloud from 2009 to the present, identifying its main contributions and limitations. We described the main steps that someone should follow to publish data streams in a manner that anyone can use them with a minimal understanding of the data details and we suggested the most suitable tool to use at every step. Based on the limitations of the current approaches, we are developing a stream publishing framework to cover the gaps. Also, we discussed the ongoing challenges and with the aim of cope some of them we propose the following directions of future work:

- (1) Conclude the implementation of the framework that is being developed.
- (2) Develop a NoSQL system as a compelling alternative to distributed and native RDF stores for simple workloads. Considering its strengths, the very large user base that it has, and the fact that there is still ample room for query optimization techniques, we are confident that NoSQL databases will present an ever growing opportunity to store and manage RDF data in the LOD cloud.

REFERENCES

- ANICIC, D. AND FODOR, P. EP-SPARQL: a unified language for event processing and stream reasoning. *Proceedings of the 20th conference on World wide web*, 2011.
- BARBIERI, D., BRAGA, D., AND CERI, S. C-SPARQL: SPARQL for continuous querying. *Proceedings of the 18th international conference on World wide web* (c), 2009.
- BARBIERI, D. AND VALLE, E. D. A proposal for publishing data streams as linked data-a position paper, 2010.
- BARNAGHI, P., PRESSER, M., AND MOESSNER, K. Publishing linked sensor data. In *CEUR Workshop Proceedings: Proceedings of the 3rd International Workshop on Semantic Sensor Networks (SSN), Organised in conjunction with the International Semantic Web Conference*. Vol. 668, 2010.
- BARRASA, J., CORCHO, O., AND GÓMEZ-PÉREZ, A. R2O, an Extensible and Semantically based Database-to-Ontology Mapping Language. In *In Proceedings of the 2nd Workshop on Semantic Web and Databases(SWDB2004)*. Springer, pp. 1069–1070, 2004.
- BERNERS-LEE, T. Linked Data - Design Issues, 2006.
- BIZER, C. D2R Map: A Database to RDF Mapping Language. In *12th World Wide Web Conference*. Budapest, Hungary, pp. 2–3, 2003.
- BRENNINKMEIJER, C. AND GALPIN, I. A semantics for a query language over sensors, streams and relations. *Sharing Data, Information and Knowledge*, 2008.
- CALBIMONTE, J., CORCHO, O., AND GRAY, A. Enabling ontology-based access to streaming data sources. *The Semantic Web ISWC* (September): 1–16, 2010.
- CALBIMONTE, J., JEUNG, H., CORCHO, O., AND ABERER, K. Semantic sensor data search in a large-scale federated sensor network. In *4th International Workshop on Semantic Sensor Networks 2011. 23 October, 2011*. Bonn, Germany., 2011.
- CARABALLO, A., JÚNIOR, N., AND NUNES, B. TRTML-A Tripleset Recommendation Tool based on Supervised Learning Algorithms. In *11th Extended Semantic Web Conference*. Anissaras, Crete, Greece, 2014.
- COMPTON, M., BARNAGHI, P., BERMUDEZ, L., GARCÍA-CASTRO, R., CORCHO, O., COX, S., GRAYBEAL, J., HAUSWIRTH, M., HENSON, C., HERZOG, A., HUANG, V., JANOWICZ, K., KELSEY, W. D., LE PHUOC, D., LEFORT, L., LEGGIERI, M., NEUHAUS, H., NIKOLOV, A., PAGE, K., PASSANT, A., SHETH, A., AND TAYLOR, K. The SSN ontology of the W3C semantic sensor network incubator group. *Web Semantics: Science, Services and Agents on the World Wide Web* vol. 17, pp. 25–32, Dec., 2012.
- COMPTON, M., HENSON, C., AND NEUHAUS, H. A Survey of the Semantic Specification of Sensors. *SSN*, 2009.
- CONSORTIUM, W. W. W. R2RML: RDB to RDF mapping language, 2012.
- HARTH, A., KNOBLOCK, C., AND STADTMÜLLER, S. On-the-fly Integration of Static and Dynamic Linked Data. In *12th International Semantic Web Conference*. Number 257641. Sydney, Australia, 2013.
- HASEMANN, H., KREMER, A., PAGEL, M., GROUP, A., AND BRAUNSCHWEIG, T. U. RDF Provisioning for the Internet of Things, 2012.
- LE-PHUOC, D., QUOC, H. N. M., PARREIRA, J. X., AND HAUSWIRTH, M. The linked sensor middleware—connecting the real world and the semantic web. *Proceedings of the Semantic Web Challenge* vol. 152, 2011.
- LOPES, G., LEME, L., NUNES, B., AND CASANOVA, M. RecLAK: Analysis and Recommendation of Interlinking Datasets. In *4th Int. Conf. on Learning Analytics and Knowledge*. Indianapolis, USA, 2014.
- LOPES, G. R., ANDR, L., LEME, P. P., NUNES, B. P., CASANOVA, M. A., AND DIETZE, S. Recommending Tripleset Interlinking. In *14th International Conference on Web Information System Engineering*. Number i. Nanjing, China, pp. 149–161, 2013.
- MORARU, A., FORTUNA, C., AND MLADENIC, D. A System for Publishing Sensor Data on the Semantic Web. *CIT. Journal of Computing and Information Technology*, 2011.
- PATNI, H., HENSON, C., AND SHETH, A. Linked sensor data. In *Collaborative Technologies and Systems (CTS), 2010 International Symposium on*. IEEE, pp. 362–370, 2010.
- PHUOC, D. AND HAUSWIRTH, M. Linked open data in sensor data mashups. In *proceedings of the 2nd International Workshop on Semantic Sensor Networks (SSN09), in conjunction with ISWC*, 2009.
- PHUOC, D. L. *A Native and Adaptive Approach for Linked Stream Data Processing*. Ph.D. thesis, National University of Ireland, 2013.
- SHETH, A., HENSON, C., AND SAHOO, S. Semantic sensor web. *Internet Computing, IEEE* 12 (4): 78–83, July, 2008.