Towards an Environmental Information System for Semantic Stream Data

Peter Wetz[1], Tuan-Dat Trinh[1], Ba-Lam Do[1], Amin Anjomshoaa[1], Elmar Kiesling[1], A Min Tjoa[1]

Abstract

The future of the earth's environmental systems will to a major extent be decided in cities as already more than 50% of the population is concentrated there. Pervasively available sensors and the data they generate can help to address pressing environmental challenges in urban areas by making crucial information available to researchers and decision-makers. However, environmental data is at present typically stored in disparate systems and formats, which inhibits reuse and recombination. Furthermore, the large amounts of environmental data that stream in continuously require novel processing approaches. So far, however, research at the intersection of environmental sciences and urban data to overcome these barriers has been scarce. To address these issues, we develop a novel framework using semantic web technologies. We apply ontological concepts and semantic stream processing technologies in order to facilitate combination, comparison, and visualization of heterogeneous data from various sources. The platform for environmental data stream analysis introduced in this paper can inform and support decision-making by non-expert users. We propose and discuss a three-step framework and outline initial results.

1. Introduction

By 2010, the share of the global population living in urban areas had surpassed 50% for the first time in history [1]. Therefore, one can argue that the future of our environment is, and will be, decided in cities. Methods and technologies developed in computer science have strong potential to improve the understanding of our environment and contribute towards solving environmental challenges [2]. Most harmful developments in urban areas are directly linked to people's behavior, which affects air and water quality, waste disposal problems, noise pollution, and the climate. One motivation for this work is to help address such negative environmental effects of urbanization with IT-based methods. [IT, for instance, can assist in analyzing combinations of traffic and air pollution data streams and thus deduce optimized traffic routing or support city planners' decision making.

Another motivation for this work is the increasingly ubiquitous presence of sensors that generate data streams. From an environmental management perspective, this can be seen as a major advantage of cities compared to rural areas. Research towards the exploitation of the data generated by such devices may lead to innovative citizen services and may ultimately help to trigger change in how we interact with the environment [3]. Means to exploit the continuously generated data, however, are still scarce. Availability of raw data can only be a first step, which has to be followed by enrichment with contextual information and careful processing to extract relevant insights.

Support efforts to provide public access to environmental information is a final key motivation for our work. The European Union (EU) Directive on public access to environmental information [4] mandates public access to and systematic distribution of environmental information through, for instance, Information and Communication Technologies (ICT). However, there are serious

[1] Vienna University of Technology, Karlsplatz 13, 1040 Vienna, Austria, Institute of Software Technology and Interactive Systems

technical barriers that inhibit citizens from readily accessing environmental information. These barriers include (i) distribution of data among different agencies and lack of a single point of access, (ii) heterogeneous storage without standardized presentation, (iii) focus on static data without accounting for the increasing importance of real-time data, and (iv) lack of embedding of data within its "context", that is, providing and utilizing additional information based on the surroundings of the data is currently not possible.

A platform that solves these challenges should exhibit characteristics including timeliness, accuracy, usability, scalability, and modularity. To the best of the authors' knowledge, there is currently no solution for the semantic integration of heterogeneous environmental data sources in (near) real-time. Taking advantage of semantic technologies in this context appears particularly useful since it facilitates both data integration and query-driven reasoning based on formalized vocabularies.

This paper outlines the architecture of a web-based platform for managing workflows and dataflows of semantically enriched environmental data. The goal for the resulting information system is to show that Semantic Web technologies – ontologies to ensure data homogeneity and RDF stream processing techniques to obtain real-time information – are suitable tools for addressing information needs in the environmental domain.

As an example imagine moving to a big city, where nonstop traffic and pollution are a given fact. It would be useful to get insights in different air pollution data, i.e., carbon monoxide, ozone, or particular matter, based on real-time values, even being able to do comparisons among each other, or combine them with other static, e.g., traffic routes, or dynamic, e.g., weather parameters, knowledge. This could help finding new insights in how our immediate environment reacts on certain events. For instance, this could lead to changes in traffic behavior or even to traffic recommendations based on deduced facts. We propose a platform that is able to tackle such questions in a novel way by exploiting and combining Semantic Web Technologies and stream processing techniques. Our vision is a Smart City system capable of measuring, sensing, analyzing and presenting the environmental "pulse" of a city, i.e., measured via characteristics such as air quality, noise pollution, water quality, and traffic information.

We tackle three main challenges in this paper, i.e., to (i) provide data in (near) real-time to support informed decisions, (ii) integrate data originating from different sources and formats, and (iii) facilitate semantic querying of the integrated stream data following Linked Data principles.

The presented work shall be seen as "work in progress" since the corresponding platform[2] is still in its early stages [5]. Nonetheless, we proof our concept by means of initial use cases based on environmental data. The remainder of this paper is organized as follows. Section 2 introduces the architecture of the platform; initial results are described in Section 3 and Section 4 discusses related work. Finally, we provide conclusions and provide an outlook on future work in Section 5.

2. Architecture

The platform for data exploitation that the contributions of this paper are built upon is called *Linked Widgets Platform*. The term *Linked Widgets* was introduced by Trinh et al. [5] to describe an extension of standard widgets [6] with a semantic model, following the Linked Data principles. This semantic model describes the input and output graph of widgets and facilitates discovery and composition of widgets into mashups. The current paper extends the architecture of the platform [7] by introducing stream processing mechanisms embodied in *Linked Streaming Widgets*.

[2] See http://linkedwidgets.org (Accessed 17 July 2014)

Our framework rests upon semantic annotations that describe the data using domain vocabularies that can be used to integrate heterogeneous environmental data. Whereas the design of this platform is domain-agnostic, we focus on real-time environmental data and the particular challenges and requirements that arise in this area in the present paper.

Figure 1 depicts the architecture, including extensions that make the platform suitable for stream data. The constituent components can be grouped into three stages, i.e., (i) data acquisition responsible for tying in polling- and streaming-based data sources, (ii) data transformation where raw data is converted into time-annotated RDF triples, and (iii) data streaming which provides streams to end-user applications.

Widgets are used to register continuous queries at the processor component. Hence, they will subscribe to RDF streams and receive corresponding data. Linked Streaming Widgets, therefore, are defined as widgets that support continuous queries with added parameters used to subscribe to data streams.

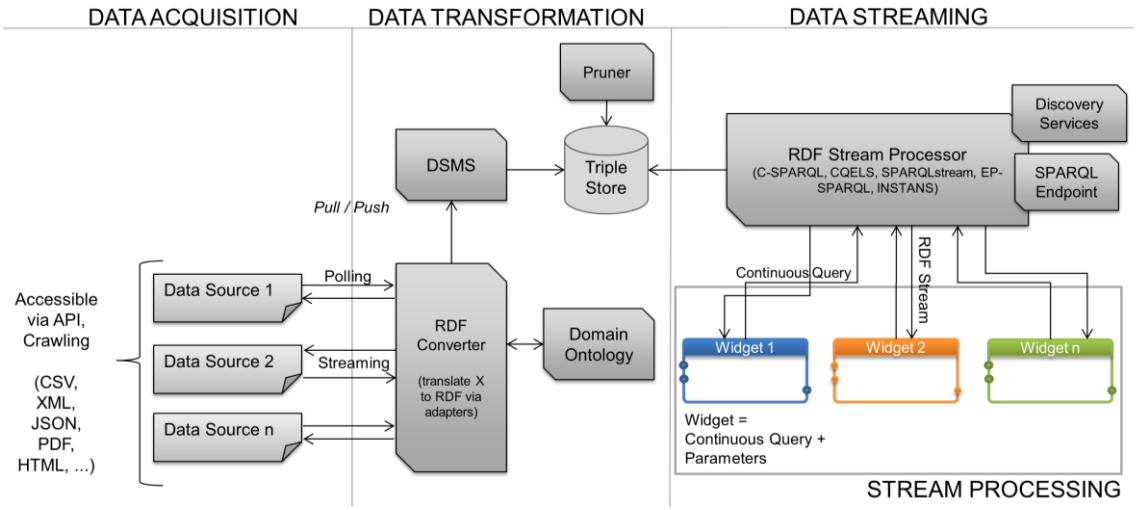

Figure 1: Linked Streaming Widgets Architecture

2.1. Data Acquisition and Data Transformation

Environmental data is available from various repositories, each providing unstructured, semi-structured, or structured data. In many cases, data is presented only on a webpage or via non-standardized interfaces. To allow for timely provision of data via our platform, such data has to be crawled on a regular basis. Data available in (semi)structured formats is more straight-forward to handle, but still needs to be converted into JSON-LD, a recent W3C recommendation [8] that we use as our internal data exchange format.

After conversion the data is fed into a Data Stream Management System (DSMS) and the triples are stored in a triple store. The RDF converter uses domain ontologies to enrich incoming data sources with semantic knowledge, which later will be utilized for features such as stream processing, or contextualized sensor discovery. The DSMS is dependent on the RDF stream processor implementation. Currently, we intend to use C-SPARQL as a stream processor which requires a semantic-aware DSMS, hence, the corresponding representation in Figure 1.

Data sources differ in type (rdf, json, xml, csv, pdf, html) and access (API, file download, manual crawling). As a means to overcome the resulting heterogeneity, ontologies have been used for many years. In the context of our proposed framework they are a valuable tool to define a

comprehensive and standardized semantic model which is a prerequisite for semantic search as well as knowledge extraction from sensor-generated data.

Furthermore, differences in number and range of observed properties as well as update frequency (varying from stream data, i.e., real-time updated data, to hourly updated data) result in large variation in the amounts of data provided which has to be taken into account when evaluating implementation candidates for the RDF converter.

2.2. Data Streaming

Stage 1 results in semantically annotated observation data, i.e., RDF streams that can be presented to end-users. In the second stage, we provide (near) real-time data to the user.

We make use of the publish-subscribe design pattern, which controls what messages are sent by entities that publish data to receiving entities [9]. In the context of the proposed framework the main advantages are (i) loosely coupled widgets can act as publishers and subscribers, and (ii) by supporting parallel operations, message caching, and routing this pattern provides the scalability needed to handle flexible stream compositions on our platform. Consequently, it solves the first step in providing environmental data streams to users by allowing clients to subscribe to data streams dynamically.

Furthermore, due to the continuity and large size of data streams, storage is a key issue. To avoid bottlenecks in subsequent procedural steps, we need to define when data becomes outdated and can be deleted. The combination of static data sources (e.g., geographic maps, point-of-interest data etc.) with dynamic data streams improves the quality of new knowledge that can be deduced. However, this blending is non-trivial and major advances still need to be made in this area.

Finally, the architecture offers flexible exploration of the data as depicted in the stream processing area of Figure 1. We achieve this by allowing users to combine small information units, i.e., widgets. This enables users to answer questions based on environmental data. Via drag and drop, these widgets can be combined into mashups. A mashup can answer information needs, e.g., display points of interest that satisfy certain air quality criteria. Widgets can be combined in many different ways leveraging the modeled semantics.

We apply stream reasoning techniques provided through SPARQL extensions, i.e., windowing functions and federation of static data with dynamic streams and combine them with a widget-based approach. One widget represents a corresponding data stream. A web-based graphical interface allows users to assemble these widgets and set parameters for their processing functions. In doing so, users will have the power to efficiently explore arbitrary data streams.

These processing widgets have encoded queries based on stream-specific criteria, e.g., time windows or aggregates (*sum*, *count*, *average*, etc.), and therefore return RDF triples that answer this query, ultimately allowing hands-on combination of data streams. Presentation widgets provide mechanisms to visualize the intended output via, for instance, maps, bar charts, line charts, pie charts, or histograms. This step covers three aspects of leveraging data streams: (i) analyzing via continuous stream queries, (ii) publishing via returning RDF graphs, and (iii) visualizing via corresponding presentation interfaces.

2.3. Semantic Modeling of Stream Data

The semantic model acts as a component which is used to annotate data streams based on domain ontologies dependent on the field the data is coming from. For the environmental domain we have already identified special vocabularies and investigated possible integration into our framework as follows.

Since ontology reuse is one important principle of the Semantic Web vision, we evaluated existing ontologies in the field of sensors and measurements. Numerous ontologies were proposed with the goal to model sensor observations. Two approaches stand out: First, the Semantic Sensor Network ontology [10], which is the result of the Semantic Sensor Networks Incubator Group at W3C. This ontology aims at a top-down approach to model whole sensor networks including sensors, observations, sites, measurement capabilities, properties, features of interest, etc. Second, the RDF Data Cube Vocabulary has been widely adopted since its promotion to a W3C Recommendation, making it an official Web Standard [11]. This vocabulary is designed for modeling observations and measurements.

In our work, we will create a new vocabulary that combines these two approaches. There is some overlap in the available concepts of both ontologies (e.g., *observations* and *properties*). These can be used to link the vocabularies. For interlinking instance data, we consider well-known domain ontologies such as the Time Ontology [12], the Basic Geo Vocabulary [13], and SWEET [14].

3. Mashup Based On Linked Streaming Widgets

Figure 2 displays an example of a mashup that uses air quality data streams, i.e., carbon monoxide and ozone, as an input. The widgets on the left hand side act as a data source. Since they are used to register a continuous query at the stream processor (see Figure 1) the necessary parameters have to be defined. The size of the window (range) and the update frequency (step) can be specified. Moreover, the user can decide whether the returned values of the query should be aggregated (min, max, average). The *Stream Merger* is needed to fuse two data streams into a single result stream that can be handled by different visualization widgets, i.e., in this case the *Line Chart* and *Google Maps Widget*. The fusion process can also be used to apply additional processing steps, e.g., transformation, aggregation, or enrichment of the incoming streams. This mashup serves as a motivating example and therefore forms the conceptual basis for our proposed framework.

By applying this approach to environmental data streams, the platform can focus more on the needs and interests of users, e.g., streams can be discovered and used based on contextual information extracted from the stream's semantics. As a result the user may discover data in his proximity, based on his/her interests, time constraints etc. and combinations of these (e.g., air quality sensor observations of the last 30 minutes within 100m of the user). Discovery based on current values, aggregates (*sum*, *median*, *mean*, *mode*, *min*, *max*, etc.) or trends (increasing, decreasing, or stagnating) is another interesting opportunity. For instance, one may be interested in analysing and comparing the pollution values (air, water, noise) near his/her appartement based on a daily or hourly basis, hence, being able to identify dynamics inherent to the data.

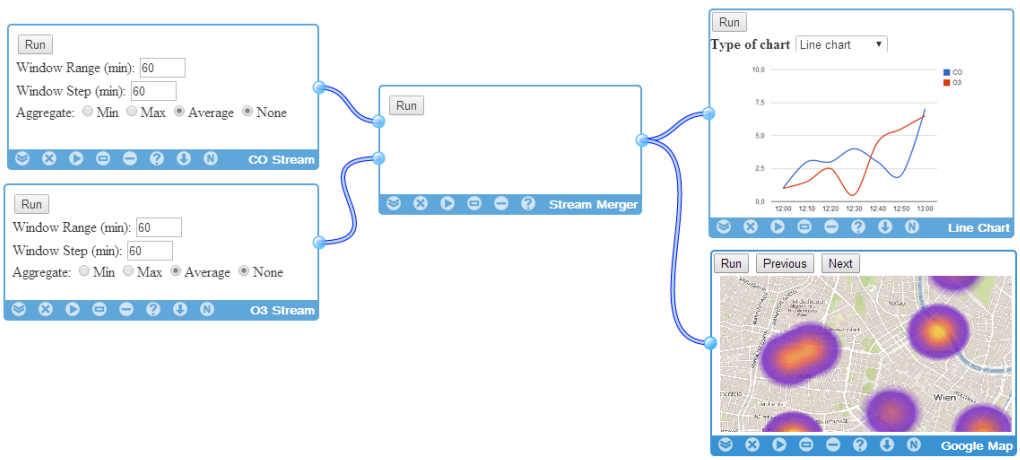

Figure 2: Example of a Mashup based on Streaming Data

4. Related Work

The amount of research in the field of semantic stream processing has been expanding rapidly in recent years. To this end, efficient means to process data streams based on semantic technologies are required in order to provide a powerful ontology-based query language utilizing continuous queries. Several approaches have recently been proposed: C-SPARQL [15], CQELS [16], SPARQLstream [17], EP-SPARQL [18], and INSTANS [19]. Moreover, a W3C RDF Stream Processing Community Group[3] has been formed to develop a definition of a common model for producing, transmitting, and continuously querying RDF streams. However, there is no complete system that supports the whole process from data acquisition to data utilization and enables flexible and efficient use of generic streams.

Balduini et al. [20] present an approach to identify events in a city leveraging a Streaming Linked Data Framework. In contrast to our work, they focus on social media, i.e., Twitter postings, as a data source. This both simplifies the semantic modeling of the data, and makes their approach dependent on geo-tagged tweets. Lastly, they do not fuse multiple social streams, but analyze a single stream at a time.

Lécué et al. [21] predict the severity of road traffic congestion using real-time heterogeneous data streams. The proposed approach is similar to ours, but focuses strongly on the traffic domain and on predictive reasoning, whereas our goal is to provide a generalized system that supports a larger spectrum of use cases.

Tallevi-Diotallevi et al. [22] present a real-time urban monitoring framework implemented for the city of Dublin. The authors extend CQELS and C-SPARQL to facilitate merging of CSV and RDF streams. Integration of other formats is not supported and explicit semantic enrichment and its subsequent utilization is not covered.

5. Conclusion and Future Work

In this paper, we propose a widget-based framework to explore environmental data streams in an urban context. We divide the approach into three stages and identify important issues that need to be addressed. These include defining a new vocabulary for environmental stream data deduced from already existing and well-adopted ontologies, and applying semantic stream processing methods to facilitate reasoning. Prototypical examples of interconnected widgets, i.e., a mashup, are explained and discussed and an architecture for a platform is outlined.

In the future, this system should serve as an open data platform for citizens of a “smart city”. The Linked Widgets Platform shall bring together both mashup developers and mashup users. For each of them, it should be as easy as possible to create, (re)use, modify, and execute available or newly created mashups. As a consequence, citizens will be enabled to interact with the available data sources, e.g., open data, linked data, tabular data, without having to worry about technical barriers such as unnecessary complexity while accessing data in different formats. New knowledge can be deduced and created by enabling creative (re)combination of available data. The vision is to provide a platform for dynamically building applications that leverage semantically enriched environmental data in a timely manner. Ultimately, this could lead to a better understanding of the environment in a local context of a city.

Future work will include implementation of a richer user interface that covers a larger number of use cases. Correspondingly, additional data sources and data input for the platform will be made available and integrated. Next to these implementation-oriented goals, we will need to find means to combine different types of data. We will also develop mechanisms to decide how long outdated

[3] http://www.w3.org/community/rsp/ (Accessed 8 July 2014)

triples will be stored and when they will be pruned. Balancing this tradeoff between being able to compare current values with historic data and the detrimental effects on performance represents an interesting challenge. Discovery Services for finding relevant sensors and data streams will be crucial as well. We will, thus, put our focus also on this aspect. In addition, as the RDF Stream Processing Group at the W3C is currently making progress towards defining a standard model for RDF stream data, we will follow this process closely.

References

[1] Wissenschaftlicher Beirat Globale Umweltveränderungen, *World in transition: a social contract for sustainability*. Berlin: WBGU, 2011.

[2] G. Huang and N. Chang, "The perspectives of environmental informatics and systems analysis," *J. Environ. Inform.*, vol. 1, no. 1, pp. 1–7, 2003.

[3] B. Resch, A. Zipf, E. Beinat, P. Breuss-Schneeweis, and M. Boher, "Towards the live city–paving the way to real-time urbanism," *Int. J. Adv. Intell. Syst.*, vol. 5, no. 3 and 4, pp. 470–482, 2012.

[4] European Union, *Directive on public access to environmental information and repealing Council Directive*. 2003.

[5] T.-D. Trinh, B.-L. Do, P. Wetz, A. Anjomshoaa, and A. M. Tjoa, "Linked Widgets-An Approach to Exploit Open Government Data," in *Proceedings of the 15th International Conference on Information Integration and Web-based Application & Services*, 2013, pp. 438–442.

[6] M. Cáceres, "Packaged Web Apps (Widgets) - Packaging and XML Configuration," *W3C Recomm.*, 2012.

[7] T.-D. Trinh, P. Wetz, B.-L. Do, A. Anjomshoaa, E. Kiesling, and A. M. Tjoa, "Open Linked Widgets Mashup Platform," in *Proceedings of the AI Mashup Challenge 2014 (ESWC Satellite Event)*, 2014, p. 9.

[8] M. Sporny, G. Kellogg, and M. Lanthaler, "JSON-LD 1.0 - A JSON based Serialization for Linked Data," *W3C Recomm.*, 2014.

[9] P. T. Eugster, P. A. Felber, R. Guerraoui, and A.-M. Kermarrec, "The many faces of publish/subscribe," *ACM Comput. Surv. CSUR*, vol. 35, no. 2, pp. 114–131, 2003.

[10] M. Compton, P. Barnaghi, L. Bermudez, R. García-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, V. Huang, K. Janowicz, W. D. Kelsey, D. Le Phuoc, L. Lefort, M. Leggieri, H. Neuhaus, A. Nikolov, K. Page, A. Passant, A. Sheth, and K. Taylor, "The SSN ontology of the W3C semantic sensor network incubator group," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 17, pp. 25–32, Dec. 2012.

[11] R. Cyganiak, D. Reynolds, and J. Tennison, "The RDF Data Cube Vocabulary," *W3C Recomm.*, 2014.

[12] J. R. Hobbs and F. Pan, "Time ontology in OWL," *W3C Work. Draft*, vol. 27, p. 133, 2006.

[13] D. Brickley, "Basic geo (WGS84 lat/long) vocabulary," *Doc. Informal Escr. En Colab.*, 2006.

[14] R. G. Raskin and M. J. Pan, "Knowledge representation in the semantic web for Earth and environmental terminology (SWEET)," *Comput. Geosci.*, vol. 31, no. 9, pp. 1119–1125, 2005.

[15] D. F. Barbieri, D. Braga, S. Ceri, E. D. Valle, and M. Grossniklaus, "Querying RDF Streams with C-SPARQL," *SIGMOD Rec*, vol. 39, no. 1, pp. 20–26, Sep. 2010.

[16] D. Le-Phuoc, M. Dao-Tran, J. X. Parreira, and M. Hauswirth, "A Native and Adaptive Approach for Unified Processing of Linked Streams and Linked Data," in *Proceedings of the 10th International Conference on The Semantic Web - Volume Part I*, Berlin, Heidelberg, 2011, pp. 370–388.

[17] J.-P. Calbimonte, O. Corcho, and A. J. G. Gray, "Enabling Ontology-based Access to Streaming Data Sources," in *Proceedings of the 9th International Semantic Web Conference on The Semantic Web - Volume Part I*, Berlin, Heidelberg, 2010, pp. 96–111.

[18] D. Anicic, P. Fodor, S. Rudolph, and N. Stojanovic, "EP-SPARQL: A Unified Language for Event Processing and Stream Reasoning," in *Proceedings of the 20th International Conference on World Wide Web*, New York, NY, USA, 2011, pp. 635–644.

[19] M. Rinne, E. Nuutila, and S. Törmä, "INSTANS: High-Performance Event Processing with Standard RDF and SPARQL.," in *International Semantic Web Conference (Posters & Demos)*, 2012, vol. 914.

[20] M. Balduini, E. Della Valle, D. Dell'Aglio, M. Tsytsarau, T. Palpanas, and C. Confalonieri, "Social listening of city scale events using the streaming linked data framework," in *The Semantic Web–ISWC 2013*, Berlin, Heidelberg: Springer, 2013, pp. 1–16.

[21] F. Lécué, R. Tucker, V. Bicer, P. Tommasi, S. Tallevi-Diotallevi, and M. L. Sbodio, "Predicting Severity of Road Traffic Congestion Using Semantic Web Technologies," in *Proceedings of the 11th Extended Semantic Web Conference*, 2014, pp. 611–627.

[22] S. Tallevi-Diotallevi, S. Kotoulas, L. Foschini, F. Lécué, and A. Corradi, "Real-Time Urban Monitoring in Dublin Using Semantic and Stream Technologies," in *Proceedings of the 13th International Conference on The Semantic Web*, 2013, pp. 178–194.