VoCaLS: Describing Streams on the Web

Riccardo Tommasini[1,a], Yehia Abo Sedira[1,b], Daniele Dell'Aglio[2], Marco Balduini[1,a], Muhammad Intizar Ali[3], Danh Le Phuoc[4], Emanuele Della Valle[1,a] Jean-Paul Calbimonte[5]

[1]Politecnico di Milano, DEIB, Milan, Italy
[a]{name.lastname}@polimi.it | [b]yehiamohamed.abosedera@mail.polimi.it
[2]University of Zurich, Zurich, Switzerland dellaglio@ifi.uzh.ch
[3]Insight Center for Data Analytics, Galway, Ireland ali.intizar@insight-centre.org
[4]Technichal University of Berlin, Berlin, Germany danh.lephuoc@tu-berlin.de
[5]University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland
jean-paul.calbimonte@hevs.ch

1 Introduction & Motivation

The interest in exploring stream publication and consumption mechanisms on the Web has recently gained attention [1,2], leveraging the progress in Stream Reasoning systems and approaches. Systems that consume streams for processing (e.g., reasoning, filtering, learning, event detection) require standards for interchanging data about the streams, including endpoint information, processing capabilities, data structure, pull and push retrieval options, etc. Although previous efforts partially tackled these problems in the past, there is still no general agreement on a shared set of principles and vocabularies for streaming data catalogs, as it is the case with *static* Linked Data.

This paper presents the highlights of the **Vo**cabulary for **Ca**taloging and **L**inking **S**treams and streaming services on the web (VoCaLS[1]). This work is a complement to the full VoCaLS paper [3], focused on the reuse, dissemination, and adoption activities related to this vocabulary. VoCaLS includes concepts related not only to the publication of streams but also the consumption and processing, regardless of implementations details and design choices of different RDF Stream Processing (RSP) and Stream Reasoning systems and languages. This vocabulary constitutes a key step towards the long-term goal of allowing Web-centered interactions among RDF Stream processing services. VoCaLS has been engineered as a collaborative effort, following the discussions and results of the work of the W3C RSP Community Group[2]. The vocabulary has been made openly available through a permanent URI, it has been submitted to the Linked Open Vocabularies (LOV) repository, it is published under a CC-BY 4.0 license, and its documentation is made available through the Widoco toolset[3].

[1] VoCaLS URI: `https://w3id.org/rsp/vocals#`
[2] `https://www.w3.org/community/rsp/`
[3] Widoco: `https://doi.org/10.5281/zenodo.591294`

2 Use-Cases & Requirements

Several use-cases motivate the design and the adoption of a vocabulary for describing streams and streaming services [1,2].

The adoption of a shared vocabulary would (i) allow decentralized & automated *discovery of streaming data publishers and consumers* at Web scale. Moreover, (ii) it would to support interactions between RDF Stream Processing (RSP) engines on the Web, standardizing the communication between them and, thus, enabling *service discovery and query federation.* Finally, *Experimentation and Empirical Research* would benefit from cataloging available streams, profiling the engine features, and tracking the provenance of the experiments.

From the aforementioned use-cases we identified the following challenges:
Publication & discovery. A *stream description* should characterize the contents of a (RDF) stream and describe the capabilities of the stream source. Moreover, a *streaming service description* should describe available endpoints from which streams can be accessed/processed/generated.
Access & processing. It is crucial to describe the capabilities of streaming services, such as stream processing engines and reasoners, in terms of their features. Moreover, it is important to allow the selection of steam partitions and windows, which can be dumped, transmitted or filtered.
Provenance & Licensing. It is required to allow tracking the transformations that involve streaming data, and those that occur on the streams, as well as contracts that regulate data access by actors involved in such transformations.

VoCaLS addresses these challenges, and in fact complies with the following requirements, which were elicited during the design phase of the vocabulary, as detailed in [3]. In summary, such vocabulary must: (i) enable the description of streams, i.e. content, relevant statistics, and the license of use; (ii) enable the description of streaming services, i.e., characterizing their capabilities, their APIs, and the license of use; (iii) enable historical stream processing/analysis and replay, i.e., allowing stream storage and dumping of stream samples; (iv) enable provenance tracking at any level, i.e., characterizing stream (a) creation, (b) publication, and (c) storage; but also denoting manipulation and management concerning to existing theoretical frameworks; (v) tame velocity for streaming data management, i.e., prioritize push-based content provisioning to pull-based one, and encouraging the adoption of an active stream processing paradigm; (vi) tame variety for streaming data management, i.e., do not bind the specification to any domain specific vocabulary, and to any specific data models, e.g., RDF Streams.

3 The VoCaLS Vocabulary of Linked Streams

The vocabulary is organized in three modules: VoCaLS Core, which describes the core elements of the vocabulary, VoCaLS Service Description, which describes RDF stream service descriptions, and VoCaLS Provenance, focused on

streaming data transformation and manipulation. We will introduce each module separately, along with illustrative examples.

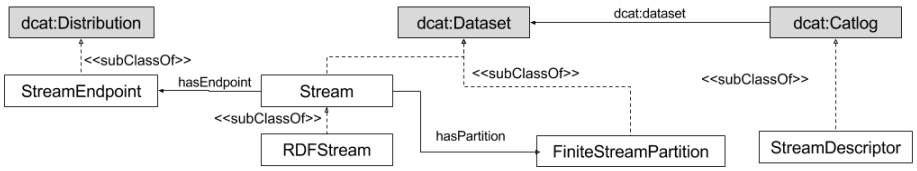

Fig. 1. VoCaLS Core module

Core Vocabulary: VoCaLS Core concepts are based on an extension of DCAT to represent streams on the Web. As depicted in Figure 1, the model introduces the basic abstractions to represent streams. A (i) `vocals:StreamDescriptor` is a document accessible via HTTP that holds metadata about the stream and its contents. A (ii) `vocals:Stream` represents a Web stream, i.e., an unbounded sequence of time-varying data elements that might be findable and accessible on the Web, and which can be consumed via a (iii) `vocals:StreamEndpoint`. Finally, a (iv) `vocals:FiniteStreamPartition` is a portion of the stream available for regular Linked Data services to access and process its content.

Streaming Service Description: VoCaLS Service Description focuses on metadata related to streaming services and their capabilities, enabling consumers to discover and select services suitable to their needs. The `vsd:StreamingService` is an abstraction to represent a service that deals data streams of any type. Continuous query engines, stream reasoners, and RDF stream publishers are valid examples. Three classes of RDF streaming services were identified, although others could be added if needed:
(i) `vsd:CatalogService`, a service that may provide metadata about streams, their content, query endpoints and more. (ii) `vsd:PublishingService`, which represents a service that publishes RDF streams, possibly following a Linked Data compliant scheme, and (iii) `vsd:ProcessingService`, which models a stream processing service that performs any kind of transformation on streaming data, e.g. querying, reasoning, filtering.

Stream Transformation Provenance: VoCaLS Provenance module focuses on tracking the provenance of stream processing services, i.e., tracing the consequences of operations performed over the streams. The module defines four main classes: (i) `vprov:R2ROperator` refers to operators that produce RDF mappings (relations) from other RDF mappings. (ii) `vprov:R2SOperator` represents operators that produce a stream from a relation. (iii) `vprov:S2ROperator` refers to operators that produce relations from streams, e.g., windowing. Finally, (iv) `vprov:S2SOperator` allows describing operators that produce a stream from another stream.

4 Discussion

Dataset description vocabularies (e.g. DCAT, DCterms, VoID) were designed primarily with static and stored (linked) data in mind, and provide metadata

descriptions for any sort of datasets published on the Web. Nevertheless, as stated before they do not allow describing Web streams and streaming services.

On the other hand, VoCaLS is a vocabulary designed for describing streams, streaming services, and it includes the capability of describing stream transformations: the operations that detail how streaming data is generated or processed. Previous attempts to cover this gap are VoIS [2] and WeSP [1], although they have several limitations regarding scope, quality, and coverage of the requirements detailed earlier in this work. These two early attempts have been used as the basis for VoCaLS, which emerged by taking the lessons learned. VoCaLS is a generic resource that can, and should, be combined with domain-specific vocabularies. The design of VoCaLS has followed a community-driven approach, starting from the W3C RSP Community group results, and a requirement analysis described in [3]. Last but not least, VoCaLS has been published following well-principled practices for the publication of the vocabulary, including the set up of permanent URIs, the availability of full open documentation using Widoco, the availability of sources in Github[4], and its inclusion in the LOV repository.

Road Map: Regarding the adoption and sustainability plans for VoCaLS, several steps have been taken in this direction. Given that The establishment of a common vocabulary is one of the main goals of the W3C RSP Community Group, we have started the process of elevating this vocabulary as an official Group Note. The adoption and support from the authors, as a relevant part of this community, will contribute positively to this endeavor. Another important goal is to foster the adoption of VoCaLS within relevant communities. For this purpose we initiated the creation of a catalog of streams descriptions[5]. Moreover, we developed a simple utility[6] to support the annotation of new streams. Finally, in order to lead by example, we have launched the integration of VoCaLS within relevant services and software available for the RSP community: the RSP Services, RSPLab, and TripleWave.

References

1. Dell'Aglio, D., Le Phuoc, D., Le-Tuan, A., Ali, M.I., Calbimonte, J.P.: On a web of data streams. In: ISWC DeSemWeb (2017)
2. Sedira, Y.A., Tommasini, R., Della Valle, E.: Towards vois: a vocabulary of interlinked streams. In: ISWC DeSemWeb (2017)
3. Sedira, Y.A., Tommasini, R., DellAglio, D., Balduini, M., Ali, M.I., Le Phuoc, D., Della Valle, E., Calbimonte, J.P.: Describing a web of streams. In: ISWC (2018)

[4] `https://github.com/ysedira/vocals`
[5] `https://github.com/ysedira/vocals/tree/master/catalog`
[6] `https://github.com/ysedira/stream-annotation-tool`