Bootstrapping the Publication of Linked Data Streams

Riccardo Tommasini[1], Mohamed Ragab[2], Alessandro Falcetta[1],
Emanuele Della Valle[1], Sherif Sakr[2]

[1]Politecnico di Milano, DEIB, Milan, Italy
[2]University of Tartu, DataSystem Group, Estonia

Abstract. Data Velocity reached the Web. New protocols and APIs (e.g. WebSockets, and EventSource) are emerging, and the Web of Data is also evolving to tame *Velocity* without neglecting *Variety*. The RDF Stream Processing (RSP) community is actively addressing these challenges by proposing continuous query languages and working prototypes. Nevertheless, the problem of Streaming Linked Data publication is still an open challenge. In this paper, we present the first attempt to tackle this challenge by introducing a set of guidelines to publish streaming linked data by reusing existing resources such as `TripleWave`, `R2RML/RML`, `VoCaLS`, and `RSP-QL`. We design a publication life-cycle that follows the W3C best practices. Besides, we present an example for publishing the Global Database of Events, Language and Tone (`GDELT`) as a Streaming Linked Data resource. We open-sourced the code of our resource and made it available for public use.

Keywords: RDF Streams, Streaming Linked Data, RDF Stream Processing, Stream Reasoning

1 Introduction

[1] A new generation of Web applications shows the need to process data reactively and continuously [3]. This challenge, known as *Data Velocity* in the Big Data domain, is now critical for the Web too [8]. The RDF Stream Processing (RSP) tries to process heterogeneous data streams that come from complex domains *on-the-fly*. Indeed, the RSP community[2] proposed the a data model, (i.e., RDF Streams) a query model (i.e., `RSP-QL`) and several resources for *processing*, and *publishing* Web streams [2, 4, 9].

The Global Database of Events, Language & Tone (GDELT) project[3] is a family of vast and heterogeneous Web Streams that is considered as the largest open-access Spatio-temporal archive for human society. Its Global Knowledge Graph spans more than 215 years, and connects people, organizations, locations, all over the world. In particular, GDELT complex from a complex domain. It

[2] `https://www.w3.org/community/rsp/`

[3] `www.gdeltproject.org`

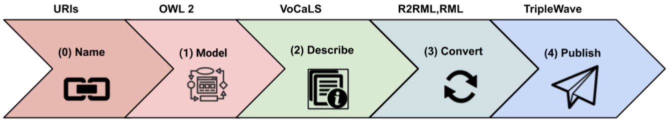

Fig. 1: Streaming Linked Data Publication Lifecycle.

captures themes, images, and emotions into a single holistic global network. GDELT data can be accessed via Google Big Query or via a number of APIs that run pre-defined analyses. Therefore, interested researches are forced to either run their analysis via the pay-per-use service or stick with the provided APIs.

GDELT Project provisions, every 15 minutes, many TSV Web Streams: the **Event Stream** makes use of the dyadic CAMEO format, describing the two actors participating in each event, and the action they perform; the **Mention Stream** records every mention of an event in the Event Stream over time, along with the timestamp the article was published; the **GKG Stream** connects people, organizations, locations, events, and more across the planet. In this paper, we present our approach for consuming and analyzing the shared GDELT TSV Streams. In particular, we provide a set of methodological guidelines that explain how to employ existing RSP resources – `RML`, `TripleWave`, `VoCaLS` – to publish and consume Streaming Linked Data. The guidelines and an extended version of the documents included in this paper are available at `http://gdelt.stream`.

2 Publishing GDELT Streams as Streaming Linked Data

In this section, we present our methodological guidelines to publish GDELT, an example of RDF streams, as Streaming Linked Data. Figure 1 depicts the steps of our publication life-cycle that has been designed following the W3C Best practices[4] and the guidelines presented by Hyland et al. [7].

The **Step (0) Name Things with (HTTP) URIs** aims of designing (HTTP) URIs that identify the relevant resources according to the Linked Data principles [7], and W3C best practices for URI design[5]. We designed URIs with the following base `http://gdelt.stream/`, distinguishing between base vocabulary (*vocab*), CAMEO (*onto/cameo*), instances (*ist*) and time instants (*time*).

The **Step (1) Model the Streams Domain** aims of (i) Understanding and capturing the domain knowledge into an ontological model, and reuse existing authoritative vocabularies. (ii) Identifying related resources, and (iii) Formulating canonical information needs [7]. This process requires collecting and reviewing applications, documentations and analyzing sample data. The data of GDELT is streamed from a multitude of news sources. The extraction process follows Natural Language Processing techniques and makes use of the Conflict and Mediation Event Observations (CAMEO) Ontology [6].

We collected all the information regarding the schemas of the streams, and we studied the CAMEO documentations. Then we designed a set of OWL2 ontologies describing the GDELT domain. The `CAMEO` ontology is a coding scheme

[4] `https://www.w3.org/TR/ld-bp/`

[5] `https://www.w3.org/TR/cooluris/#cooluris`

which is designed for the study of third-party mediation in international disputes. It contains a hierarchical coding-scheme for dealing with sub-state actors, event types, and an extensive taxonomy for religious groups and ethnic groups. For CAMEO, we model event and actors types, into a comprehensive hierarchy.

The **Step (2) Describe the Stream** aims for providing human-readable and machine-readable representations of the streams [7] as commonly done for datasets [1]. To this extent, Tommasini et al. proposed a vocabulary based on DCAT[6] named `VoCaLS` [10].

```
1  <> a vocals:StreamDescriptor ; dcat:dataset :eventstream
       .
2  :eventstream a vocals:Stream ;
3   dcat:title "GDELT Event Stream"^^xsd:string ;
4   dcat:publisher <http://www.streamreasoning.org> ;
5   dcat:description "GDELT Events Stream"^^xsd:string ;
6   vocals:windowType vocals:logicalTumbling ;
7   vocals:windowSize "PT15M"^^xsd:duration ;
8   vocals:hasEndpoint [
9     a vocals:StreamEndpoint ;
10    dcat:license <https://cc.org/licenses/by-nc/4.0/> ;
11    dcat:format frmt:JSON-LD;
12    dcat:accessURL "ws://examples:8080/events" ] .
```

Listing 1.1: VoCaLS GDELT Stream description.

We used VoCaLS to describe GDELT streams. Listing 1.1 shows an example of description for the GDELT Event Stream. In `VoCaLS`, a Web Stream is represented using *vocals:Stream*, i.e., an unbounded sequence of data items that might be accessible on the Web. A *vocals:StreamDescriptor*, i.e, a HTTP-accessible document that contains stream metadata. Finally, the stream content can be consumed via a *vocals:StreamEndpoint*, that refer to actual sources using *dcat:accessURL* property. GDELT does not use a license format, thus we include a license that is compliant with the terms of use. We also linked ontologies and mappings using `rdfs:seeAlso`. Since the stream is published regularly as 15 minutes batch, we include metadata about the rate, i.e., at lines 6-7 `vocals:windowType` and `vocals:windowSize`.

```
<GEM> a rr:TriplesMap ; rml:logicalSource  <source> ;
 rr:subjectMap [
  rr:template "http://gdelt.stream/ist/{GLOBALEVENTID}";
  rr:class gdelt:Event;
  rr:graphMap [
   rr:template "http://gdelt.stream/time/{DATEADDED}"]];
 rr:predicateObjectMap [ rr:predicate rdf:type;
  rr:objectMap [
   rr:template "http://gdelt.stream/cameo/{EventCode}"
        ;]];
 rr:predicateObjectMap [
  rr:predicate gdelt:actor;
  rr:objectMap [ rr:parentTriplesMap <Actor1TM> ]];...
```

Listing 1.2: Subset of GDELT RML Mapping.

The **Step (3) Convert to RDF Stream** aims of enabling RDF data provisioning so that data could be enriched with domain knowledge [7] from Step (1). The conversion to RDF can occur either automatically or via expert modeling [7]. Approaches based on mapping languages decouple conversion and modelling, and support the conversion process using a formal language like `R2RML` [7] for relational data, and `RML` for logical sources [5] sharing document-based formats (e.g., JSON), or semi-structured formats (e.g., CSV). Similarly, Web streams are often not shared as RDF Streams and needs to be converted. In practice, GDELT content is not natively RDF. Therefore, we need to set up a conversion mechanism. For data conversion, we followed the RML approach where CSV data is converted using its mappings. Listing 1.2 shows a sample of

[6] `https://www.w3.org/TR/vocab-dcat/`

[7] `https://www.w3.org/2001/sw/rdb2rdf/r2rml/`

such R2RML mapping. Notably, we used `rr:class` to assign the `gdelt:Event` type to the data at line 4. Moreover, we assign the CAMEO type using `rdf:type`.

The **Step (4) Publish The RDF Stream** aims for serving the data to the audience of interest. Datasets – opportunely described with contextual vocabularies [1] – are either added to the Linked Open Data (LOD) Cloud, shared using REST APIs, or exposed via SPARQL endpoints. Critical aspects of this step are licensing, audit, and access control [7].

The unbounded nature of Streaming Linked Data makes it impossible to directly add a data stream to the LOD Cloud [10, 9]. However, using an RSP Engine that focuses on RDF Stream provisioning like `TripleWave` [9], one can publish a stream describing using `VoCaLS`. `TripleWave` relies on Barbieri et al's vision [2] of serving a "static" named graph called `S-Graph` via REST API, to identify and describe the streams. Moreover, they envisioned that the stream elements of an RDF Graph are called `I-Graph`s. We shared the VoCALS description files like the one in Listing 1.1 as S-GRAPH via REST APIs. Since the concept of event is first-class, we opted for a graph-based stream data model.

3 Conclusion

This paper presents a first attempt to publish streaming Linked Data following W3C's best practices [7]. Some real-world Web streams from the GDELT project are available as RDF Stream, described using VoCaLS, and accessible at `http://gdelt.stream`. Future work include publishing other Web streams towards a catalog of Linked Streams.

References

1. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets. In: Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009, Madrid, Spain, April 20, 2009. (2009)
2. Barbieri, D.F., Della Valle, E.: A proposal for publishing data streams as linked data - A position paper. In: Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010 (2010)
3. Della Valle, E., Dell'Aglio, D., Margara, A.: Taming velocity and variety simultaneously in big data with stream reasoning: tutorial. In: DEBS (2016)
4. DellAglio, D., Della Valle, E., van Harmelen, F., Bernstein, A.: Stream reasoning: A survey and outlook. Data Science 1(1-2), 59–83 (2017)
5. Dimou, A., Sande, M.V., Slepicka, J., Szekely, P.A., Mannens, E., Knoblock, C.A., de Walle, R.V.: Mapping hierarchical sources into RDF using the RML mapping language. In: International Conference on Semantic Computing (2014)
6. Gerner, D.J., Schrodt, P.A., Yilmaz, O., Abu-Jabr, R.: Conflict and mediation event observations (cameo). International Studies Association, New Orleans (2002)
7. Hyland, B., Wood, D.: The joy of data-a cookbook for publishing linked government data on the web. In: Linking government data, pp. 3–26. Springer (2011)
8. Margara, A., Urbani, J., van Harmelen, F., Bal, H.E.: Streaming the web: Reasoning over dynamic data. J. Web Sem. 25, 24–44 (2014)
9. Mauri, A., Calbimonte, J., Dell'Aglio, D., Balduini, M., Brambilla, M., Della Valle, E., Aberer, K.: Triplewave: Spreading RDF streams on the web. In: ISWC (2016)
10. Tommasini, R., Sedira, Y.A., Dell'Aglio, D., Balduini, M., Ali, M.I., Phuoc, D.L., Della Valle, E., Calbimonte, J.: Vocals: Vocabulary and catalog of linked streams