FraPPE: a vocabulary to represent heterogeneous spatio-temporal data to support visual analytics.

Marco Balduini, Emanuele Della Valle

DEIB, Politecnico of Milano, Milano, Italy
marco.balduini@polimi.it, emanuele.dellavalle@polimi.it

Abstract. Nowadays, we are witnessing a rapid increase of spatio-temporal data that permeates different aspects of our everyday life such as mobile geolocation services and geo-located weather sensors. This big amount of data needs innovative analytics techniques to ease correlation and comparison operations. Visual Analytics is often advocated as a doable solution thanks to its ability to enable users to directly obtain insights that support the understanding of the data. However, the grand challenge is to offer to visual analytics software an integrated view on top of multi-source, geo-located, time-varying data. The abstractions described in the FraPPE ontology address this challenge by exploiting classical image processing concepts (i.e. Pixel and Frame), a consolidated geographical data model (i.e. GeoSparql) and a time/event vocabulary (i.e. Time and Event ontologies). FraPPE was originally developed to represent telecommunication and social media data in an unified way and it is evaluated modeling the dataset made available by ACM DEBS 2015 Grand Challenge.

1 Introduction

Nowadays, we are witnessing a rapid increase of sources exposing geo-located time-varying data such as social media, mobile telecommunication, taxis, etc. Making sense of those data sets typically requires to compare and correlate them. Visual Analytics – the science of analytical reasoning facilitated by interactive visual interfaces [**?**] – is often advocated as an effective solution for those tasks.

For instance, Figure 1 illustrates a real case of visual analytics for a general audience[1] where: a grid of 6x3 cells is overlaid to a city street map, green circles represent the number of tweets posted in a time interval from each cell, and the fill colour opacity value of each cell is mapped to the number of mobile calls from each cell. As shown in [**?**], people without specific expertise in data analytic can easily spot the cells where the two signals are correlated.

However, data is not often ready for visual analytics. Usually, geo-located time-varying data of this type has first to be aggregated over time and space.

[1] Interested readers are invited to view `https://youtu.be/MOBie09NHxM`

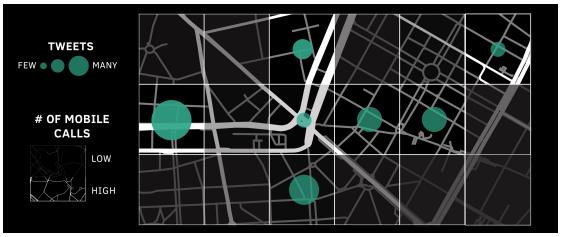

Fig. 1. A real-world example of visual analytics of two heterogenous datasets.

FraPPE is a vocabulary designed exactly for this purpose following Methontology [?] guidelines. It exploits classical image processing concepts (FRAme and Pixel) – familiar to designers of visual analytics solutions – as well as common sense concepts (Place and Event) using a consolidated geographical data model (GeoSparql [?]) and a time/event vocabulary (Time [?] and Event ontologies [?]).

The basic building blocks of FraPPE were originally developed to represent, in an homogenous way, heterogeneous data streams for the CitySensing[2] installation proposed to the public of Milano Design Week 2014. Some of its concepts (i.e., dividing the physical space in cells using a grid and linking time-varying data to cells) were used to publish the dataset of Telecom Italia Big Data Challenge 2014[3]. In this paper, we present the first formal version of this vocabulary (Section 2) and we evaluate it (Section 3) by assessing its compliance to Tom Gruber's principles (i.e., clarity, coherence, minimal encoding bias, minimal ontological commitment, extendibility) modelling the dataset of ACM DEBS 2015 Grand Challenge[4].

FraPPE vocabulary is published in Linked Open Vocabularies[5] while community discussion, issue tracking and advancement are handled via github[6]. Those resources are maintenined and sustained on the long term by the Stream Reasoning research group of Politecnico di Milano. The data of ACM DEBS 2015 Grand Challenge modelled in FraPPE can be queried using a SPARQL endpoint[7] whose content is described with VoID [?] machine processable metadata[8].

2 FraPPE

The overall idea of FraPPE is depicted in Figure 2. A portion of the physical world is illustrated using a map in the top-right side of the figure. A GRID, which, in this example, is made of 4 CELLs, sits between the physical world and the

[2] http://citysensing.fuorisalone.it/
[3] https://dandelion.eu/datamine/open-big-data/
[4] http://www.debs2015.org/call-grand-challenge.html
[5] http://lov.okfn.org/dataset/lov/vocabs/frappe
[6] https://github.com/streamreasoning/FraPPE
[7] http://www.streamreasoning.com/datasets/debs2015/
[8] http://streamreasoning.org/datasets/debs2015/void.rdf

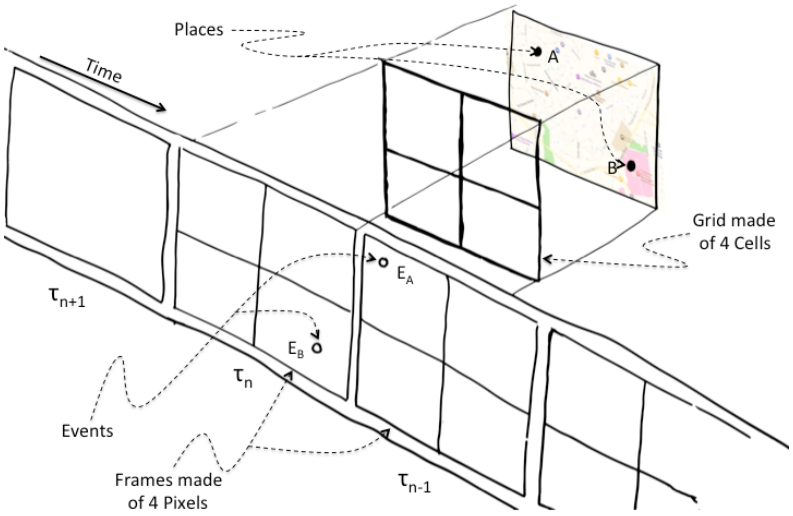

Fig. 2. An high-level view of FraPPE including 3 FRAmes made of 4 Pixels containing Places where Events happens.

film that CAPTURES a FRAME per time-interval. The frame being captured in the current time interval τ_n is directly in front of the grid. The previous frame, captured at time τ_{n-1}, is on its right. The next empty frame, which will be captured at time τ_{n+1} is on the left. The two captured frames both have 4 PIXELs (one for each cell). The physical world contains two PLACEs (e.g. A the start and B the end points of a journey). The frame captured at τ_{n+1} contains a pixel that accounts for the EVENT E_A occurred in A at τ_{n+1} (e.g., the pick up of some good). In a similar manner, the just captured frame accounts for the event E_B occurred in B at τ_n (e.g., the drop-off of some good).

More formally, FraPPE ontology is organised in three interconnected parts: the geographical part, the time-varying one and the provenance one (see Figure 3). PLACE, CELL and GRID belong to the geographical part and reuse geosparql vocabulary [**?**]. They are geosparql `Feature`s whose default geometry are respectively a `point`, a `surface` and a `multisurface`. EVENT, PIXEL and FRAME are in the time-varying part. The EVENT concept is borrowed from the Event ontology [**?**]. The provenance part includes the activities CAPTURE and SYNTHETIZE (see also Figure 4) and reuses the PROV Ontology [**?**] (PROV-O).

An EVENT has a LOCATION in a PLACE that is `sfWithin`[9] a CELL – the basic spatial unit of aggregation of information in FraPPE – which, in turn, is `sfWithin` a GRID.

A PIXEL is the time-varying representation of a CELL. It is the only element in the conceptual model that carries information through the HASVALUE data

[9] `sfWithin` refers to the `within` relationship defined in the Simple Features standard jointly issued by Open Geospatial Consortium and ISO.

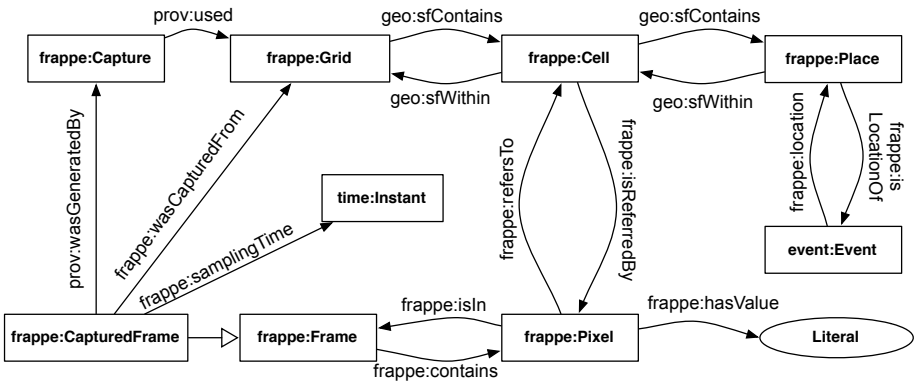

Fig. 3. The core terms of FraPPE vocabulary, which reuse the geosparql vocabulary, the PROV Ontology and the Event Ontology.

property. As in image processing, this value represent a measure of intensity of some phenomena in the real world. For instance, it can represent the number of micro-posts posted in a given time interval within a certain Cell. Each Pixel refersTo a single Cell, contrariwise a Cell could be referredBy many different Pixels that captures different information associate to the same Cell, e.g., the already mentioned number of micro-posts, but also the number of mobile phone calls or the number of good pick-ups.

Similarly, a Frame is the time-varying counterpart of a Grid. It is a single complete *picture* in a series forming a *movie*. FraPPE distinguishes between two types of frames: the CapturedFrames and the SyntheticFrames (see Figure 4). A CapturedFrame contains a Pixel for every Cell in the Grid it wasCapturedFrom. Different Frames represent different images of the observed phenomena at the same samplingTime, e.g., a frame captured the volume of the social activity while another one captured the volume of the mobile phone calls at 12.00.

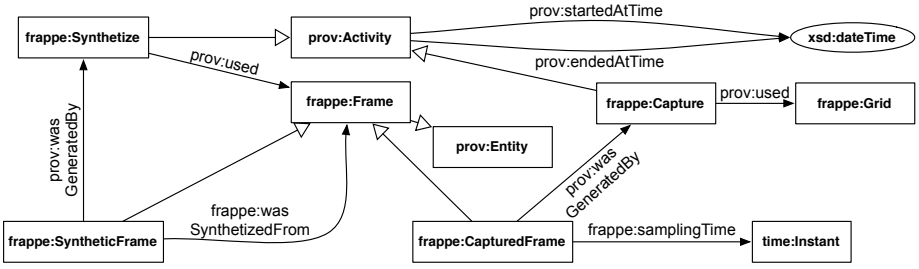

Fig. 4. The part of FraPPE that reuses the provenance ontology.

Figure 4 provides more details on the provenance part of FraPPE . The CapturedFrame and the SyntheticFrame are specializations of Frame which is an `Entity` in PROV-O. Also Grid is an `Entity`. This is because FraPPE proposes the ternary relationships Capture and Synthetize as specialisations of the relationship `Activity` of PROV-O.

This allows to model that a CapturedFrame `wasGeneratedBy` a Capture `Activity` `startedAtTime` τ_i and `endedAtTime` τ_j that `used` a given Grid. The object property wasCapturedFrom is the result of the chaining of those two `wasGeneratedBy` and `used` object properties. Moreover, the value of the samplingTime data property, which describes the CapturedFrame, is the one assigned to the `startedAtTime` data property that describes the captured activity.

Similarly, a SyntheticFrame `wasGeneratedBy` by a Synthetize `Activity` that `used` one or more Frames. The idea is to derive a frame from one or more others. The synthetize operation can be a filter applied to the values of the pixels, or an aggregation of values of pixels across frames or the difference between the values associated to the pixels of two different frames. For a fully fledge algebra of some of the operations we intend to model, we refer interested readers to [**?**].

For instance, in our work on CitySensing [**?**], we captured for 2 months a frame every 15 minutes associating the value of each pixel to the volume of mobile phone calls in the 10,000 cells we divided Milan into. In this way, we captured 96 frames per day. Then, we synthetized 96 frames (one for each slot the day is divided into) associating to the value of each pixel a Gaussian distribution ϕ_{avg,std^2} where avg is the average and std is the standard deviation of the values associated to the respective pixels in the captured frames. Each pixel is, thus, associated with a statistical model of the volume of mobile phone calls. With these models, given a pixel in the frame captured at 12.00, whose value is v, we can compute an anomaly index to associate to the value of a pixel in a new synthetic frame using the formula $2\phi_{avg,std^2}(v) - 1$. A value of that pixel close to 1 (or -1) indicates an extraordinary higher (lower) volume compared to the usual activity in that slot from 12.00 to 12.15 in the associated cell of Milan.

3 Evaluation

In order to evaluate FraPPE, we check if it observes the five principles of Tom Gruber [**?**]: clarity, coherence, minimal encoding bias, minimal ontological commitment and extendibility.

FraPPE observes the *clarity* principle because all definitions are documented in natural language (see version of FraPPE on github[6]). The terms proposed in FraPPE are: (i) common terms in spatial-related vocabularies (e.g., Place, Cell, Grid); (ii) well known terms of the image processing domain (e.g., Pixel, Frame, Capture, or Sythetize); and (iii) terms defined in other ontologies (e.g., Event, Instant, Entity, or Activity). Moreover, they are independent of the social and telecommunication domains, for which FraPPE was originally defined.

Listing 1. Fraction of the model representing ACM DEBS Grand Challenge 2015 Data

```
@prefix frGrid: <http://streamreasoning.org/debsGC/Grids/> .
@prefix frCell: <http://streamreasoning.org/debsGC/Cells/> .
@prefix frPixel: <http://streamreasoning.org/debsGC/Pixels/> .
@prefix frPlace: http://streamreasoning.org/debsGC/Places/:> .
@prefix frEvent: <http://streamreasoning.org/debsGC/Events/> .
@prefix frFrame: <http://streamreasoning.org/debsGC/Frames/> .
@prefix frCapture: <http://streamreasoning.org/debsGC/Captures/> .

frGrid:Grid_1
    gs:sfContains frCell:Cell_1, frCell:Cell_2 .

frCell:Cell_1
    a fr:Cell ;
    rdfs:label "39460"^^xsd:long ;
    fr:isReferredBy frPixel:1356995100000_39460 ;
    gs:sfContains frPlace:A ;
    gs:sfWithin frGrid:Grid_1 .

frPlace:A
    a sf:Point ;
    fr:isLocationOf frEvent:E_B ;
    gs:asWKT "POINT( 40.715008 -73.96244 )"^^gs:wktLiteral ;
    gs:sfWithin frCell:Cell_1 .

frEvent:E_A
    a fr4d:PickUpEvent ; a event:Event ;
    event:time [ a time:Instant ;
            time:inXSDDateTime "2013-01-01T00:00:00"^^xsd:dateTime ] ;
    fr:location frPlace:A> ;
    fr4d:hackLicense "E7750A37CAB07D0DFF0AF7E3573AC141"^^xsd:string ;
    fr4d:medallion "07290D3599E7A0D62097A346EFCC1FB5"^^xsd:string .

frEvent:E_B
    a fr4d:DropOffEvent ; a event:Event ;
    event:time [ a time:Instant ;
            time:inXSDDateTime "2013-01-01T00:02:00"^^xsd:dateTime ] ;
    fr:location frPlace:B ;
    fr4d:connected frEvent:E_A ;
    fr4d:fareAmount "3.5"^^xsd:double ;
    fr4d:mtaTax "5.0"^^xsd:double ;
    fr4d:paymentType "CSH"^^xsd:string ;
    fr4d:surcharge "5.0"^^xsd:double ;
    fr4d:totalAmount "4.5"^^xsd:double ;
    fr4d:tripDistance "0.44"^^xsd:long ;
    fr4d:tripTime "120"^^xsd:long .

frPixel:1356995100000_39460 a fr:Pixel ;
    fr:isIn frFrame:1356995100000 ;
    fr:refers frCell:Cell_1 .

frFrame:1356995100000
    a fr:CapturedFrame ;
    fr:contains frPixel:1356995100000_39460, frPixel:1356995100000_39461 ;
    fr:samplingTime [ a time:Instant ;
            time:inXSDDateTime "2013-01-01T00:05:00"^^xsd:dateTime ];
    fr:wasCapturedFrom frGrid:Grid_1 ;
    prov:wasGeneratedBy frCapture:1356995100000 .
```

Indeed, FraPPE terms can be used for other domains as demonstrated in publishing data for the Telecom Italia Big Data Challenge[3].

FraPPE is *coherent*, i.e., all FraPPE inferences at T-box level are consistent with the definitions and in modelling A-boxes containing social, telecommunication, environment, traffic, and energy consumption data, we never inferred inconsistent or meaningless data.

FraPPE has a *minimal encoding bias* because it is encoded in OWL2-QL. Indeed, it uses only subclass axioms, property domain, property range and inverse object properties. We explicitly avoided adding cardinality restrictions, because in CitySensing [**?**] we use FraPPE to integrate data following an ontology-based data access approach.

FraPPE requires a *minimal ontological commitment*, meaning that, as Tom Gruber recommended, FraPPE makes as few claims as possible about the geo-located time-varying data being modelled allowing who uses FraPPE to specialise and instantiate it as needed.

Last but not least, we tested in details that FraPPE is *extendable* by modelling the dataset made available by ACM DEBS 2015 Grand Challenge[4]. The challenge proposes a taxi route analysis scenario based on a grid of 150x150 Kms with cells of 500x500 m. A stream of data represents the route of a taxi rides in terms of: (i) taxi description, (ii) pick-up and drop-off information (e.g., geographical coordinates of the place and time of the event), and (iii) ride information (e.g., tip, payment type and total amount). In the Listing 1, we report a subset of the information representing a single taxi ride in FraPPE . The pick-up EVENT represents the start of the ride and contains the taxi id. The drop-off EVENT represents the end of the trip and it is connected to all the information about the ride. The fragment models the geographical part of the ride using two PLACEs within two different CELLs of a single GRID. Moreover, it models the time varying-part of the ride using two EVENTs captured in two PIXELs of a single FRAME along with the provenance part through the CAPTURE activity. Indeed, we reuse all FraPPE concepts, we specialise EVENT in `PickUpEvent` and `DropOffEvent`, and we add attributes (e.g., `tripTime`, and `totalAmount`) and an object property (i.e., `connected`) specific of the taxi ride domain.

Synthetic frames are also important in the challenge. One of the problems, assigned to the challengers, asks to compute the profitable cells for a given time interval. We wrote a SPARQL query that computes the answer; this is the SYTHETIZE activity that `used` CAPTUREDFRAMEs of the type illustrated in Listing 1 to construct a SYTHETICFRAME where the values of the PIXELs are associated with the profitability of the CELLs they refer to.

4 Conclusions

In this paper, we propose FraPPE, a novel vocabulary that fills in the gap between low-level time-varying geo-located data and the high-level needs of map-centric visual analytics. Vocabularies to publish the low-level data exist, e.g., geosparql vocabulary [**?**], event ontology [**?**] or time ontology [**?**]. FraPPE reuses them. The high-level (oriented to visual analytics) part is missing, but the terms, which we choose for FraPPE vocabulary, are largely used among practitioners. The genesis of FraPPE terms can be found in the Social Pixel approach, proposed in [**?**], and in common practices used in geo-spatial knowledge discovery and data mining domain, e.g., [**?**] where the authors discuss on the optimal size of CELLs in GRIDs used for the analysis of human mobility.

FraPPE has an high potential. We demonstrated it applying FraPPE to city data integration within CitySensing [**?**]. A preliminary version of FraPPE was used in 2014 to publish open the data of Telecom Italia Big Data Challenge that covers the telecommunication, social, environmental, and energy domains. In this paper, we further exemplify such a potential by publishing in RDF the dataset of the ACM DEBS Grand Challenge 2015.

In designing FraPPE , we followed Tom Gruber's principles. Moreover, FraPPE is described using machine processable metadata (i.e., `label`, `creator`, `issued`, `versionInfo`, `priorVersion`, `license`, and `imports`). The ACM DEBS Grand Challenge dataset is accessible via SPARQL[7] and described using VoID[8].

As future works, we want to investigate how to improve the modelling of the values associated to the pixels. We started an investigation on ontologies of the units of measurement. Our best candidate, at this moment, is [**?**], because it allows also to model dimensionless quantities such as the anomaly index. Moreover, we want to investigate an effective approach to link a Grid to Where On Earth Identifiers[10] in order to define a unique ID for every possible Cell on earth. Last, but not least, we want to foster the adoption of FraPPE by publishing more datasets; our intention is to start from the datasets release as open data within the Telecom Italia Big Open Data Challenge 2014.

References

1. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets. In: LDOW 2009 (2009), `http://ceur-ws.org/Vol-538/ldow2009_paper20.pdf`
2. Balduini, M., Della Valle, E., Azzi, M., Larcher, R., Antonelli, F., Ciuccarelli, P.: Citysensing: visual story telling of city-scale events by fusing social media streams and call data records captured at places and events. IEEE MultiMedia 22(3), to appear (2015)
3. Battle, R., Kolas, D.: Geosparql: enabling a geospatial semantic web. Semantic Web Journal 3(4), 355–370 (2011)
4. Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: PROV-O: The PROV Ontology. Tech. rep., W3C (2012)
5. Coscia, M., Rinzivillo, S., Giannotti, F., Pedreschi, D.: Optimal spatial resolution for the analysis of human mobility. In: ASONAM 2012. pp. 248–252 (2012)
6. Fernández-López, M., Gómez-Pérez, A., Juristo, N.: Methontology: from ontological art towards ontological engineering (1997)
7. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing? Int. J. Hum.-Comput. Stud. 43(5-6), 907–928 (1995)
8. Hobbs, J.R., Pan, F.: Time Ontology in OWL (September 2006)
9. Raimond, Y., Abdallah, S.: The event ontology (2007), http://motools.sf.net/event
10. Rijgersberg, H., van Assem, M., Top, J.L.: Ontology of units of measure and related concepts. Semantic Web 4(1), 3–13 (2013)
11. Singh, V.K., Gao, M., Jain, R.: Social pixels: genesis and evaluation. In: ICM 2010. pp. 481–490 (2010)
12. Thomas, J.J., Cook, K.A.: Illuminating the Path: The Research and Development Agenda for Visual Analytics. National Visualization and Analytics Ctr (2005)

[10] `https://developer.yahoo.com/geo/geoplanet/guide/concepts.html`