An Ontology Design Pattern for Microblog Entries

Cogan Shimizu and Michelle Cheatham

Data Semantics Laboratory, Wright State University, Dayton, OH, USA

Abstract. Due to the exponential growth of the Internet of Things and use of Social Media Platforms, observers have an unprecedented level of detailed information available on the behavior of communities. However, due to the highly heterogeneous nature and the immense volume of the data, a composite view is difficult to generate. Such a composite view would be exceptionally useful in the realms of insider threat detection, after-action forensics, and hazardous situation detection and avoidance. The Semantic Web, via ontology modeling, offers a powerful tool for fusing the disparate data sources and formats. To this end, we have created an ontology design pattern (ODP) for the modeling of a simple microblog entry. This ODP is intended to fit within an ecosystem for fusing social media, support advanced visualization, and provide a preliminary framework for trust assessment.

1 Motivation & Scope

In recent years, access to data has become increasingly trivial as Social Media Platforms and the Internet of Things (IoT) continue to grow. However, important latent or implicit information runs the risk of obfuscation simply by the sheer volume of collected data. Further, the data is presented and accessed via highly disparate vectors (e.g. microblog entries, visual media, and geotagged textual data). Thus, it is increasingly necessary to identify and develop methods for seamless fusion and visualization of information extracted from heterogeneous social media data.

Such methods are especially important for obtaining an accurate and comprehensive view of a crisis theater or battlespace (e.g. formulating a "Common Operating Picture"¹). For these use cases, it is also important to take into account the provenance and trustworthiness of the acquired data and for any conclusions drawn from such data. To support the fusion of such heterogeneous data and the capture of its metadata, we will build an ecosystem of ontology design patterns [6]. ODPs enable sophisticated visualizations that leverage the inherent concept hierarchy, such as models displaying varying levels of granularity and interconnectedness. Figure 1, provides two examples of possible visualization methods that the microblog entry (MBE) will help support. We are currently

¹ A Common Operating Picture is a single identical display of relevant operational information on materiel shared by more than one Command. This term is frequently

investigating other visualizations in collaboration with domain experts from the United States Air Force. In this paper, we describe a pattern for a MBE as an entry point into developing the ecosystem.

The MBE pattern is important for a number of reasons. First, microblog entries are representative of a fairly large subset of publicly available social media data. For example, Twitter² the popular, public-facing microblogging platform, allows a Tweet's payload to contain text, hyperlinks, images, or video. The entries may also be geotagged and may explicitly refer to other users. Additionally, there are many existing datasets that capture Tweets during natural disasters and humanitarian crises (e.g. CrisisLex³).

By definition and intent, microtext⁴ is simple; its model is relatively straightforward and requires little of the complexity that OWL brings to the table. Regardless, it is important to note that this pattern is a fundamental building block of the intended ODP ecosystem. However, due to its simplicity, it is relatively straightforward to fit with many existing patterns. Specifically, we foresee easy integration with the ModifiedHazardousSituation Design Pattern [4] and ReportingEvent [7]. As the ecosystem matures, we also foresee including existing patterns regarding maps, climate, and public infrastructure.

Finally, the MBE pattern has some components that allow for interesting interaction: spatiotemporal extent and author trustworthiness. Spatiotemporal extent of information is of particular interest to the modeling community as there are still many open questions on its handling. However, it is an integral part of any sort of response or intelligence operation. In a perfect world, we could assume that any author neither seeks to mislead nor propagate lies. However, in light of recent events, as well as the ODP's relevance to crisis and operational intelligence management, it is necessary to include a component for the trustworthiness of an author. Thus, the model for the microblog entry seeks to answer, at least, the following competency questions. Due to the strong emphasis on geospatial and temporal components of the fused data, we assume that these queries will be executed using geoSPARQL⁵.

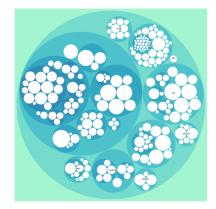
- 1. Who is the author of entry x?
- 2. What are all the entries authored by y?
- 3. What entries from time A to time B originate from region of interest C with radius D?
- 4. What is the trust value v for author y?
- 5. What is the trust value v for entry x?
- 6. What entries from authors with a trust value greater than v originate from a region of C with radius D?
- 7. What entries relate to topic T?

² https://twitter.com

³ http://crisislex.org/

⁴ Microtext is any sufficiently short parcel of information in natural language. An MBE is an instance of microtext.

⁵ http://www.opengeospatial.org/standards/geosparql



(a) A Circle Packing visualization generated by D3⁶. Smaller circles are related to the superimposed circle via subsumption and proximity in the same level of circle denotes a short semantic distance.



(b) A standard view of geographic information: pins on a map background. This visualization can be updated in real-time and allows the user to see incoming data.

Fig. 1: Both visualizations will utilize the MBE pattern at the most granular level (i.e. smallest circles and map pins).

Microtext is a valuable resource in the Semantic Web Community, as evidenced by [2, 9, 10, 8]. However, to our knowledge this is the first attempt at modeling an MBE as an entity, instead of only modeling extracted information.

The rest of the paper is organized as follows. Section 2 will address the design decisions in the structure of the pattern and accompanying axioms. Section 3 provides a motivating example and interaction with real data. Section 4 addresses future work and collaborations.

2 Pattern Overview

This pattern was directly informed by the competency questions in the preceding section; the competency questions are fairly straightforward and have a one to one correspondence with the concepts in the pattern. As such, the microblog entry pattern must capture both the entry's payload and its provenance. In addition, it must capture any information extracted from the payload and analysis of the author, such as answers to the questions: "To what is the microblog entry referring?" or "How trusted is the author by their peers?"

We will discuss the main design aspects of this pattern by referring its class diagram as depicted in Figure 2. Yellow boxes indicate datatypes, light blue boxes with dashed borders indicate external patterns. Purple is used for external

⁶ Circle Packing is an arrangement of circles on a surface so that all circles touch one another. D3 is a powerful JavaScript library used for generating visualizations.

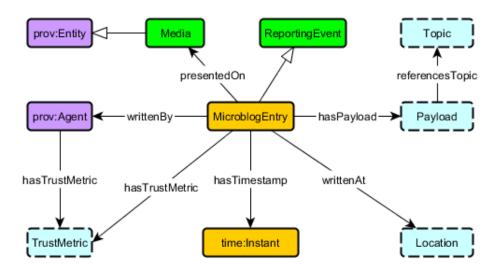


Fig. 2: A graphical representation of the microblog entry design pattern. Yellow boxes indicate datatypes, light blue boxes with dashed borders indicate external patterns. Purple is used for external classes belonging to PROV-O [5]. Green is used for external classes belonging to [7]. White arrowheads represent the owl:SubclassOf relation.

classes belonging to PROV-O [5]. Green depicts external classes belonging to [7]. White arrowheads represent the owl:SubclassOf relation.

By indicating several of the classes as "external," we intend to convey that the models for said classes are not indicative of the functionality of the MicroblogEntry pattern. For example, in our implementation⁷ the light blue boxes are currently wrappers for datatypes. However, it is not hard to imagine increasingly complex models for each class. Below, we will discuss our implementation and future iterations. We will consider the pattern in the context of our use-case: event detection during a crisis. Furthermore, we assume that any microblog entry populating the ontology occurs within the time-frame and are shown to be relevant to the crisis situation.

MicroblogEntry The MicroblogEntry is the core class. Here, we will describe a few limitations placed upon its relations.

⁷ The OWL file can be found at https://raw.githubusercontent.com/ cogan-shimizu-wsu/MicroblogEntryOWL/master/MicroblogEntry.owl

$MicroblogEntry \sqsubseteq = 1 hasPayload.Payload$	(1)	
---	----	---	--

 $\mathsf{MicroblogEntry} \sqsubseteq = 1 \mathsf{hasAuthor}. \mathsf{Author}$ (2)

 $\mathsf{MicroblogEntry} \sqsubseteq \leq 1 \mathsf{hasLocation}. \mathsf{Location} \tag{3}$

- 1. A MicroblogEntry may only have one Payload.
- $2.\ A$ MicroblogEntry may only have one Author.
- 3. A MicroblogEntry might not have a location attached to it.

ReportingEvent The ReportingEvent pattern is documented in [7]. This established pattern provides for a lot of interplay with MicroblogEntry, as well as providing structure for how information is shared.

As ReportingEvent is itself a subclass of Situation, it will be reasonably straightforward to integrate the ModifiedHazardousSituation [4] pattern to the MicroblogEntry. Additionally, ReportingEvent provides a framework for connecting the "report" to an ActualEvent; thus, along with Topic, ground the MicroblogEntry in reality. Finally, the fact that a ReportingEvent *isBasedOn* a Source, provides us a vehicle for capturing the fact that a MicroblogEntry has been re-Tweeted or shared (without modification).

Media The Media class allows us to represent the platform on which the MicroblogEntry was posted. In the case of our example in the next section, this would be Twitter. However, it is also conceivable that Media may represent CNN, Fox News, BBC, and so on. Obviously, these establishments are fairly complex in their own right.

Media is also drawn from [7], though is largely left for others to implement. Monitoring different Media will be very important in our use case scenario, especially when considering the TrustMetric for provenance and author. To this point, it seems reasonable to expect the trustworthiness of the platform and corporation to effect the trustworthiness of the reported data.

Payload The Payload is the content of the MicroblogEntry. In Figure 3, this is the content in Box 2. For the general pattern, we opted to leave this as an external pattern due to the expected heterogeneity of MBEs of different platforms and even high variance of content on the same platform. That is, Twitter allows for many different payloads: text, hyperlinks, images, and videos. Facebook, on the other hand, offers a superset of content types and no length restriction on text payloads.

In addition, we see the Payload playing a large role in defining how MBEs will interact with each other. In the case of Tweets, a Tweet may be "Retweeted," thus embedding a Tweet inside of a Payload. Furthermore, a Payload may "mention" another user or author. Our next steps will include ways to more accurately model these relationships between Authors, Payloads, and MicroblogEntries. For our initial implementation, as our test sets do not include Tweets with pictures or hyperlinks, Payload wraps an xsd:string. Additionally, relevant MicroblogEntries must have a relevant Payload. That is, the Payload must refer to some Topic relevant to the crisis situation.

Topic In some cases, it may make sense to have **Topic** include a targeted list of terms from a controlled vocabulary. Or, instead, to have the **Topic** act as a category. For example, in [3], Tweets were partitioned into the following categories: affected individuals, infrastructures and utilities, donations and volunteer, caution and advice, sympathy and emotional support, useful information and unknown.

Our implementation currently wraps an xsd:string. This allows us to dynamically generate a Topic as Tweets are encountered. As the intended ODP ecosystem matures, it is conceivable that this Topic sub-pattern will be more fully fleshed out, allowing for more interesting interaction between MicroblogEntries referencing the same Topic.

Location There are many methods for representing location, e.g. the POI:Place [1] pattern or using WellKnownText (WKT) from OpenGIS, among others. To promote reusability, we do not constrain the top-level pattern to use one or another. In our implementation, however, we opted to use a WKT literal for simplicity's sake. In the future, we expect to be able to augment this part of the model by including relevant descriptors, such as the name of the location taken from a gazetteer.

TrustMetric The TrustMetric sub-pattern has the potential to be the most complex due to its far reaching effects on the interplay between Author, Payload, and Media. In addition, the actual metric for trust will need its own provenance and uncertainty measures. Until the system is actually implemented, it will be difficult to completely model. Thus, in our implementation, we assume we are getting a value between 0 and 1 from some black-box system. As such, we wrap xsd:double.

3 Example Triples

Figure 3 shows an example Tweet. The relevant data that will be extracted has been boxed in red.

kast:CarAccident	## Extracted from Box 2
rdf:type	t:Topic;
t:hasName	"Car Accident"^^xsd:string;
kast:Evacuation	## Extracted from Box 2



Fig. 3: An example Tweet with extracted data highlighted in red. Note, this example does not have a geolocation.

rdf:type t:hasName	t:Topic; "Evacuation"^^xsd:string;
kast:examplepayload rdf:type kast:hasvalue	<pre>## Extracted from Box 2 pl:Payload; "There is a car accident on 4th and Main. Be careful out there! #evac"^^xsd:string;</pre>
<pre>kast:referencesTopic .</pre>	kast:CarAccident, kast:Evacuation;
kast:cogantm rdf:type tm:hasValue	<pre>## Note here that there are two trust metrics. tm:TrustMetric; .99^^xsd:double;</pre>
kast:mbetm rdf:type tm:hasValue	<pre>## As trust in author is distinct from trust in the MBE. tm:TrustMetric; .89^^xsd:double;</pre>
<pre>kast:CoganShimizu a prov:Person, prov:Agen foaf:givenName kast:hasTrustMetric .</pre>	<pre>## Extracted from Box 1 t; "Cogan Shimizu"^^xsd:string; kast:cogantm;</pre>
kast:Twitter rdf:type	<pre>pz:Media, prov:Entity;</pre>

```
kast:examplets  ## Extracted from Box 3
rdf:type  time:Instant;
time:inXSDDateTimeStamp "2017-07-12T10:01:00-5:00"^^xsd:dateTimeStamp;
```

And finally,

```
kast:exampletweetrdf:typekast:MicroblogEntry, pz:ReportingEvent;kast:hasPayloadkast:examplepayload;kast:writtenBykast:CoganShimizu;kast:presentedonkast:Twitter;kast:hasTrustMetrickast:mbetm;kast:kastTimestampkast:examplets;
```

4 Conclusions and Future Work

The Microblog Entry Ontology Design Pattern is a useful model for a very commonplace structure, especially as the amount of social media data available for inspection continues to increase. The potential applications of this pattern are widespread, from determining public sentiment, measuring affect, or investigating community formation and evolution on social media networks.

The Microblog Entry pattern is foundational. On its own, it is not particularly remarkable. However, in the ecosystem it plays a fundamental role. In similar systems, it is analogous to entity extraction. Knowing the entities in play is important, but ultimately provides only a small facet of a crisis situation. The Microblog Entry pattern serves a similar role. It provides the threads to weave a more comprehensive picture. At this time, the pattern heavily relies on many external patterns, though many of them can be implemented as simple wrappers for datatypes. Future work will be focused on developing the ecosystem of ODPs for building a Common Operating Picture for a crisis situation. We will also investigate how the different visualizations can be effected by the trust metric. As the work progresses, we will be working closely with domain experts in the United States Air Force.

Acknowledgement. The authors acknowledge support by the Dayton Area Graduate Studies Institute (DAGSI) and input from Vincent Schmidt, Ph.D.

References

 A. Alves, B. Antunes, F. C. Pereira, and C. Bento. Semantic enrichment of places: Ontology learning from web. Int. J. Know.-Based Intell. Eng. Syst., 13(1):19–30, Jan. 2009.

- S. P. Bhatt, H. Purohit, A. Hampton, V. Shalin, A. Sheth, and J. Flach. Assisting coordination during crisis: A domain ontology based approach to infer resource needs from tweets. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 297–298, New York, NY, USA, 2014. ACM.
- G. Burel, H. Saif, M. Fernandez, and H. Alani. On semantics and deep learning for event detection in crisis situations. 2017. Available from http://semdeep.iiia. csic.es/files/SemDeep-17_paper_5.pdf on September 6, 2017.
- M. Cheatham, H. Ferguson, C. Vardeman, and C. Shimizu. Modified hazardous situation odp. 2017. Available from http://www.michellecheatham.com/files/ modification-hazardous-situation.pdf on September 6, 2017.
- P. Groth and L. Moreau, editors. PROV-Overview: An Overview of the PROV Family of Documents. W3C Working Group Note 30 April 2013, 2013.
- P. Hitzler, A. Gangemi, K. Janowicz, A. Krisnadhi, and V. Presutti, editors. Ontology Engineering with Ontology Design Patterns: Foundations and Applications. Studies on the Semantic Web. IOS Press, Amsterdam/AKA Verlag, Heidelberg, 2016.
- E. Kowalczuk and A. Lawrynowicz. The reporting event ontology design pattern and its extension to report news events. 2017. Available from http://ontologydesignpatterns.org/wiki/images/a/ac/WOP2016_paper_ 18.pdf on September 6, 2017.
- M. B. Lazreg, M. Goodwin, and O. Granmo. Information abstraction from crises related tweets using recurrent neural network. In L. S. Iliadis and I. Maglogiannis, editors, Artificial Intelligence Applications and Innovations - 12th IFIP WG 12.5 International Conference and Workshops, AIAI 2016, Thessaloniki, Greece, September 16-18, 2016, Proceedings, volume 475 of IFIP Advances in Information and Communication Technology, pages 441–452. Springer, 2016.
- R. Nithish, S. Sabarish, M. N. Kishen, A. M. Abirami, and A. Askarunisa. An ontology based sentiment analysis for mobile products using tweets. In 2013 Fifth International Conference on Advanced Computing (ICoAC), pages 342–347, Dec 2013.
- P. Thakor and S. Sasi. Ontology-based sentiment analysis process for social media content. *Procedia Computer Science*, 53:199 – 207, 2015. INNS Conference on Big Data 2015 Program San Francisco, CA, USA 8-10 August 2015.