# Using On-the-Fly Pattern Transformation to Serve Multi-Faceted Event Metadata

Stasinos Konstantopoulos

Institute of Informatics and Telecommunications,
NCSR 'Demokritos', Greece
konstant@iit.demokritos.gr

**Abstract.** In this paper we present an extension of DOLCE UltraLite and Event Model F developed for the SYNC3 repository, storing semantic information about news content and the world events that such content documents. In the SYNC3 ontology we introduce a conceptualization of web documents and also propose an alternative mereological hierarchy for text document that the one in DOLCE. Finally, we introduce the idea of using inference to provide multi-faceted querying access to the data.

## 1 Introduction

In this paper we present the ontology developed for the SYNC3 data store, and we introduce the idea of using inference to provide multi-faceted querying access to the data. The SYNC3 data store manages and serves relations between world events and the news content that documents them, as well as metadata such as events' thematic category, location and time, participating named entities, related events, and the sentiment expressed in news content towards them.

The SYNC3 ontology extends the DOLCE UltraLite and Event Model F models, and is the schema of a large-scale triple store holding automatically extracted data, generated at a rate of roughly 40 million triples per month. In Section 2 we discuss some key decisions and points of divergence from these foundations, and motivate the decision to diverge.

In Section 3 we proceed to present a novel approach to multi-faceted querying that enables using different (previously coordinated) ontological schemas to query the same data. Our approach uses inference to dynamically generate data in different facets, avoiding the reduplication of data at such a large scale. This approach is discussed in Section 4, where future work is also outlined.

## 2 The SYNC3 Ontology

The SYNC3 domain is that of news and events described in news articles and blog posts, so that the concepts of a text *document* and of a news-worthy *event* reported in it are prominently situated in the SYNC3 model.

We shall not delve into the details of the linguistic processing pipeline of SYNC3 [1]; it suffices to say that at the end of this processing, the following information about documents and events has been extracted:

- Document metadata, including title, date of publication, and source.
- A breakdown of documents into *segments*, each comprising consecutive syntactic elements of the document which document the same event. Besides extracting events, the *sentiment* (if any) expressed in each segment towards the event is also extracted.
- The resolution of the *abstract domain entity* that each concrete term, pronoun or other anaphora in the document refers to.
- The geographical and temporal grounding of an event, as well as a numerical valuation of the level of participation of domain entities in each event.

### 2.1 Extending DUL/F

The SYNC3 ontology[1] is based on DOLCE+DnS UltraLite (DUL),[2] a modular foundational ontology which is the Description Logic-compatible subset of the DOLCE ontology [2].

DUL is a lightweight foundational ontology for modelling both physical and social contexts, extensions of which have been successfully applied in several domains. Most DOLCE modules have been ported to DUL, but particularly pertinent to SYNC3 are:

- *Descriptions and Situations* (DnS), conceptualizing social entities such as relations, roles, contexts, situations, and parameters; and
- *Information Objects*[3] (IOLite), covering expressions and meaning, logical and physical documents, and reference.

Event Model F [3] extends DUL+DnS to represent the participation of agentive, temporal, spatial, and other entities in events, as well as temporal, causal, and generic correlative relationships between events. Furthermore, Event Model F supports event composition and alternative interpretations of the same event.[4]

Most pertinent to the work described here is Event Model F's *participation pattern* that links an event participation description (kinds of participation) with specific objects (participants). This approach offers Model F the flexibility to have participant instances assume different roles in different event patterns without the need to define new sub-properties of the `f:hasParticipant` relation, but rather by populating the ontology with event role instances.

The IOLite concept of `io:InformationRealization` is specialized to web content as the `sync3:DigitalDocument` concept: the class of information realizations that occupy a `sync3:WebArchive` region. A `sync3:WebArchive` is the

[1] `http://www.sync3.eu/rdf/sync3` abbreviated hereafter to `sync3:`
[2] `http://www.loa-cnr.it/ontologies/DUL.owl` abbreviated hereafter to `dul:`
[3] `http://www.loa-cnr.it/ontologies/IOLite.owl` abbreviated hereafter to `io:`
[4] `http://events.semantic-multimedia.org/ontology/2009/4/15/model.owl` abbreviated hereafter to `f:`

subclass of `dul:SpatioTemporalRegion` that specifies a particular web location at a particular time. Its instance's properties carry crawling meta-data, such as URL and time of crawling, as well as a key that identifies a specific web document from a specific crawl stored in the SYNC3 multimedia repository. Furthermore, the spatial component of `sync3:WebArchive` instances can optionally have the `sync3:startsAt` and `sync3:endsAt` properties, restricting the region to the fragment between these two token indexes.

What should be noted is the distinction between the localization of the information object and its realization: the spatio-temporal specification of information objects refers to the time and place where the object was authored, as extracted from the object itself. The `sync3:WebArchive` instances that specify the, one or more, realizations of this objects conceptualize that a concrete digital object was retrieved by the system from a certain URL at a certain time point.

### 2.2 A New Mereology of Information Objects

SYNC3 extends the IOLite `io:LinguisticObject` pattern to represent meta-data and named-entity extraction results; linguistic objects are information objects where information is expressed in natural language.

The IOLite mereological organization of `io:Text`, `io:Sentence`, `io:Phrase`, `io:Word` was deemed inappropriate for SYNC3 because its axiomatization forbids the omission of any of its levels. Since SYNC3 processing follows a bag-of-words approach, `io:Phrase` and `io:Sentence` instances are not extracted. Furthermore, named-entity recognition in SYNC3 extracts multi-word references to an entity, a significant level between `io:Phrase` and `io:Word` that is missing from IOLite. For these reasons, the SYNC3 ontology defines its own mereology of linguistic objects, comprising `sync3:Text`, `sync3:Segment`, and `sync3:DomainTerm`, linked in a mereology by the `dul:hasComponent` relation. In this model:

- a `sync3:Text` instance represents a complete document,
- a `sync3:Segment` instance represents the *maximal semantically homogeneous* fragment of a `sync3:Text` instance, such that is a *single* `dul:Entity` can fill its `dul:expresses` property . In SYNC3, this filler is a `dul:Event` so that `sync3:Segment`S represent the maximal fragments of the article such that all entities mentioned in them are participants in the same event.
- a `sync3:DomainTerm` instance represents the *minimal* `sync3:Text` fragment that has a semantics and can be linked to a `dul:Entity` via `dul:expresses`. In SYNC3, `sync3:DomainTerm` instances are (possibly multi-word) references to domain entities (persons, organizations, locations, etc.)

We believe this mereology to be not only more appropriate for the SYNC3 application, but a model that is *generally* better suited to modern information extraction systems, as well as more flexible than the rigid sentence/phrase/word hierarchy imposed by IOLite. For example, applications where full-depth syntactic analysis is used to extract the full compositional semantics of the text could represent the complete analysis as a tree of appropriate specializations the

sync3:Segment class, where the dul:Entity expressed by each is an expression that composes the semantics of the sub-segments of this segment. From this perspective, sync3:DomainTerm is the sub-class of sync3:Segment that has semantics that cannot be further decomposed but are references to semantic units in the domain of discourse.

## 3 Pattern Transformation as Inference

The SYNC3 repository is implemented within the OpenRDF Sesame framework[5] and its architecture of *Storage And Inference Layers* (SAILs). Sesame SAILs are 'stackable' components that infer implicit RDF triples from the (explicit or also implicit) data they receive from the SAIL immediately below.

### 3.1 The LODE SAIL

We have implemented a SAIL that infers data following the *Ontology for Linking Open Descriptions of Events* (LODE)[6] given Event Model F data.

Both event models annotate events with a dul:Location and a spatio-temporal dul:Region using the dul:hasLocation and dul:hasRegion properties. These do not require any transformation. More interesting is the transfer of data about event participants: The two event models are different but compatible, in that both make a distinction between the participation of dul:AgentS and the participation of other dul:ObjectS in an event, and that both use a sub-property to denote that the participation of dul:AgentS is a special kind of participation.

$$\frac{\begin{array}{l} sss\ \texttt{rdf:type}\ \texttt{f:ParticipationSituation} \\ sss\ \texttt{dul:includesEvent}\ eee \quad sss\ \texttt{dul:includesAgent}\ xxx \end{array}}{eee\ \texttt{lode:involvedAgent}\ \texttt{xxx}}$$

$$\frac{\begin{array}{l} sss\ \texttt{rdf:type}\ \texttt{f:ParticipationSituation} \\ sss\ \texttt{dul:includesEvent}\ eee \quad sss\ \texttt{dul:includesObject}\ xxx \end{array}}{eee\ \texttt{lode:involved}\ xxx}$$

However, among non-agentive participants, LODE can only model the participation of dul:Object instances, as lode:involved is restricted to range over dul:Object.

The more generic relation dul:isSettingFor between a situation and any dul:Entity instance is not transferred, even if its filler falls under one of the cases above. The rationale is that there is no axiom in DUL that forces a dul:isSettingFor relation between a situation and a dul:Object to assume the semantics of the more specific dul:includesObject property.

[5] See http://www.openrdf.org

[6] http://linkedevents.org/ontology hereafter abbreviated as lode:

In the SYNC3 ontology, news content is modelled as `dul:InformationObject` patterns, using `dul:isAbout` to link then to the `f:ParticipationSituation` that are reporting. In the LODE model this link is more direct, since instances of `dul:InformationObject` link directly to the `dul:Event` instance. The rule below infers this link:

$$\frac{\begin{array}{l} io \texttt{ rdf:type dul:InformationObject} \\ sss \texttt{ rdf:type f:ParticipationSituation} \\ io \texttt{ dul:isAbout } sss \quad sss \texttt{ dul:includesEvent } eee \end{array}}{io \texttt{ lode:illustrate } eee}$$

### 3.2 Discussion

What can be observed is that these rules do more than simply re-writing property names, as there is a significant change of perspective between the two models: in LODE the `dul:Event` instance assumes a more 'central' position in the pattern, being the only instance that is directly linked to all other instances in the pattern. Event Model F, on the other hand, is more closely adhering to the DUL foundation by extending the generic Description/Situation pattern, so that the Situation instance is the 'central' element of the pattern.

Another important point is that the transfer of data between different event models is achieved by Java code specific to DUL and LODE. In order words, the *ontology coordination* knowledge about the correspondences between two ontological schemas is encoded as Java code rather than in a knowledge representation formalism.

On the positive side, most of this code deals with model-idependent tasks such as retrieving the statements that make up the source pattern, recursively applying the transformation to their property fillers, etc. The part that maps triples can be easily generalized to read the mapping from a knowledge base that encodes knowledge about the coordination of the two schemas.

The decision to encode the model-specific knowledge in the implementations of the event interface was taken based on the absence of a stable and generally-accepted schema for representing ontology coordination knowledge; all code design decisions were taken in anticipation of such a schema that will enable the development of a generic implementation of the event interface.

## 4 Conclusions

The main contributions of this paper are the extension and deployment of the DUL and Event Model F foundational ontologies on a large-scale application in the domain of world events and the on-line news content that reports them; and the development of a novel approach to multi-faceted access to data that dynamically generates facets without the need to reduplicate information in order to serve data under a different perspective.[7]

[7] The code pertinent to DUL/Model F is published as part of the TransOnto knowledge management and transformation system, `http://transonto.sourceforge.net`

In Section 2.2 we have also identified a problem in the DUL/IOLite conceptualization of text and its fragments, where the rigidness of the proposed structure makes it inappropriate for modelling the results of modern information extraction applications. The proposed solution is both better suited for SYNC3 and similar systems, but can also be extended to models equivalent to the current DUL/IOLite conceptualization.

As RDF repositories become larger, dynamically generating alternative facets from coordinated conceptualizations of the same data will become a key enabling technology, avoiding the need to reduplicate information at a large scale. In the case of the SYNC3 system, for example, extracting roughly 38Mtriples per month, statically storing alternative facets would impose a prohibitive burden. In Section 3 we propose a novel approach for using the customized inference architecture available in many modern RDF frameworks in order to dynamically generate alternative facets.

Future research plans involve developing a vocabulary of OWL annotation properties that can provide meta-information about the ontological schema, such as identifying the 'central' instance of a pattern that links to every other instance and the property (or chain of properties) through which this instance reaches every other instance in the pattern. This will enable the development of a generic mechanism of serving facets without reference to any two particular ontologies.

## Acknowledgements

The author wishes to acknowledge the support of the European FP7-ICT project SYNC3, `http://www.sync3.eu`

## References

1. Sarris, N., Potamianos, G., Renders, J.M., Grover, C., Karstens, E., Kallipolitis, L., Tountopoulos, V., Petasis, G., Krithara, A., Gallé, M., Jacquet, G., Alex, B., Tobin, R., Bounegru, L.: A system for synergistically structuring news content from traditional media and the blogosphere. In: Proceedings of the eChallenges e-2011 Conference, Florence, 26–28 October 2011. (2011)
2. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening ontologies with DOLCE. In Gómez-Pérez, A., Benjamins, V.R., eds.: Proc. 13th Intl. Conf. on Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web (EKAW 2002), Siguenza, Spain, 1–4 Oct. 2002. LNCS 2473, Springer Verlag, Berlin/Heidelberg (2002)
3. Scherp, A., Franz, T., Saathoff, C., Staab, S.: F – a model of events based on the foundational ontology DOLCE+DnS Ultralite. In: Proc. 5th Intl Conf. on Knowledge Capture (K-CAP 2009), Redondo Beach, California, 1–4 September 2009, ACM (2009)

---

Code specific to the SYNC3 ontology is part of public SYNC3 Deliverable 4.2, September 2011. Please see `http://www.sync3.eu` or contact author.