“And the winner is...” – Representing awards on the Web of Data

Julien Plu[1,2] and Alexandre Passant[1]

[1] seevl.net, MDG Web Limited
Unit 201, Business Innovation Centre,
NUI Galway, Galway, Ireland
julien,alex@seevl.net – http://seevl.net
[2] Université Montpellier 2, Sciences et Techniques,
Place Eugène Bataillon,
34095 Montpellier Cedex 5, France
julien.plu@etud.univ-montp2.fr – http://www.univ-montp2.fr/

Abstract. Awards are a particular kind of events happening in domains as various as entertainement, sport, and even scientific conferences. Here, we present an approach to model events on the Web of Data, and to extract event information about music artists from Wikipedia categories, combining (1) data extraction using background knowledge, and (2) machine learning with the Google Prediction API.

Keywords: Awards, events, music, Wikipedia, machine learning, natural language processing, SPARQL

1 Introduction

Whether it is in music, sport or academic conferences, awards are frequently granted to people for a particular accomplishment, such as the best live performance or the best album in the case of music. Yet, there is currently no easy way to query this information on the Web of Data[2]. In this paper, or goal is to provide an infrastructure to find who won such-and-such award, and in which category, while having the results provided in a machine readable data, i.e. RDF.

While a knowledge base like Wikipedia provide this information through categories (as people can be assigned to categorie describing an award, e.g. “Academy Award winner”, the category does not explitely represent an award not its category. These Wikipedia categories can indeed include the name of the award, and sometimes the name of the category, but this information has a poor semantic, and a better structure is needed. To overcome this, we designed (1) an ontology to represent awards, and (2) a tool to extract this data from Wikipedia categories.

The remainder of this paper is organised as follows. First, we present a state of the art of ontologies represent events and, more particularly, awards. Then, we detail our *Award ontology*, from its design rationale to instances representation. In the third part, we present our approach for extracting such information from Wikipedia using wikipedia categories, etc.

2 State of the art

2.1 Generic event ontologies

An event can be described as a public gathering for the purpose of celebration, education, marketing or reunion. Events can be classified on the basis of their size, type and context. For instance, we consider the following as being events:

- *social and lifecycle* events: birthday parties, graduation days, weddings;
- *education and career* events: workshops, conferences, debates;
- *sports* events: football tournaments, Olympics Games, Roland Garros;
- *entertainment* events: music awards, concerts, festivals;
- *political* events: debates, summer schools, assembly meetings;

Hence, we consider award ceremonies as being events as they represent an important fact which brings together many people for a particular focus.

In order to model such events in a machine-readable way, ontologies are an obvious candidate. The field of event ontologies has been widely studied, and the following models can be considered to represent events on the Web of Data:

- The Event Ontology[3]: the most appropriate ontology to describe the representation of an event. It has a simple model, and has been extended in several domains, such as by the Press Association with their own event ontology[4];
- LODE: An ontology for Linking Open Descriptions of Events[5][5]: a complete event ontology, based on the previous Event Ontology and DOLCE;
- KMI ontology[6]: an ontology used for event extraction, providing very specific details, such as the event duration, sub-events, meeting organisers, etc;
- Event-F model[4]: a very detailed ontology for describing an event, and all their related properties;
- CIDOC-CRM [1]: an ontology specifically made for historical events.

While the Event Ontology and LODE provide enough details to represent events, they lack features to represent awards, for example their category (sport, music, etc.). Similarly, the KMI ontology and Event-F do not provide this support, and were too specific for our use-case. Finally, by being tailored for historical events, CIDOC-CRM was not directly appropriate neither for awards representation.

2.2 Award-related ontologies

In addition to the previous models, the following ontologies can more specifically describe awards:

- BBC sport ontology[6]: tailored for sport-related events, and awards, but not adapted for other kind of awards, such as music;

[3] http://motools.sourceforge.net/event/event.html

[4] http://data.press.net/ontology/event/

[5] http://linkedevents.org/ontology/

[6] http://www.bbc.co.uk/ontologies/sport/2011-02-17.shtml

- SWPortal ontology[3]: a perfect model to academic conferences awards to represent a conference award, but not appropriate to describe a sport award by example;
- Sport-ontology[7]: similar to the BBC sport ontology, but with a simplier model (and less specific) for describing a sport event;
- Baseball Ontology[8]: an ontology describing baseball events, but not appropriate for other domains.

3 Award Ontology

The main purpose of this ontology was how to describe an award event for a music artist. Our needs were to describe the ceremony where an artist won an award (e.g. "MTV Music Awards") as well as the category of this award (e.g. "Best song"). Not only it is useful to describe the event, but it can help to answer the following queries:

- list all artists that have been nominated and won in a category for a given ceremony;
- list all the awards and nominations of an artist on the same year;
- identify the career path of an artist, e.g. winning the "Best song" the year after winning the "Best your talent" award.

To do so, we extended the event ontology and developed a lightweight awards ontology, available at `http://seevl.net/schema/awards`, described in the following schema. The ontology classes are:

- *Award*: a subclass of *Event:Event*, representing the Award itself, as an event;
- *Ceremony*: a class representing the ceremony corresponding to an event (e.g. "MTV Music Awards");
- *Category*: a subclass of *skos:Concept*, describing a category in a ceremony (e.g. "Best song").

and its properties:

- *time*: date of the award described (e.g. "The award took place the 20/07/2012");
- *place*: place of the award described (e.g. "The award took place in Paris");
- *ceremony*: corresponding to the ceremony of an award (e.g. "The ceremony is MTV Music Awards");
- *category*: corresponding to one or many categories of an award (e.g. "The category is Best Song");
- *win*: a subproperty of *isAgentOf*, meaning that the artist won the award in a single or many categories;
- *participates*: a subproperty of *isAgentOf*, meaning that the artist is a participant in one or many categories;

[7] `http://code.google.com/p/sport-ontology/`

[8] `www.daml.org/2001/08/baseball/`

- *nominated*: a subproperty of *isAgentOf*, meaning that the artist is nominated in an award in one or many categories;

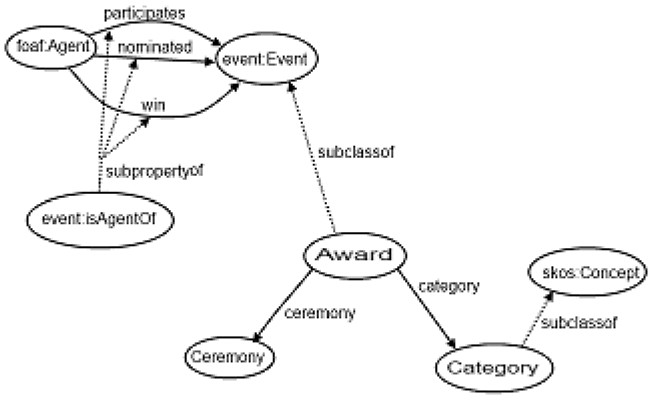

Fig. 1. award ontology schema

We extensively reused the Event Ontology to describe an event corresponding to an award, and we redefined the *isAgentOf* property which links a *foaf:Agent* to an *Event:Event* in order to identify who won, is nominated or participates in a ceremony, and more precisely at an award.

We also identified the need to organise categories in a hierarchical fashion, for instance considering that "Best album" is a top-category of both "Best world-music album" and "Best alternative rock album". We rely on SKOS to represent these hierarchies, representing *Category* as a sub-class of a *skos:Concept*[9]:

```
ex:best_album rdf:type awards:Category .

ex:best_world_music_album> rdf:type awards:Category ;
  skos:broader ex:best_album .

ex:best_alternative_rock_album rdf:type awards:Category ;
  skos:broader ex:best_album .
```

A simple complete example of award representation can be:

[9] Prefixes ommitted for space reason

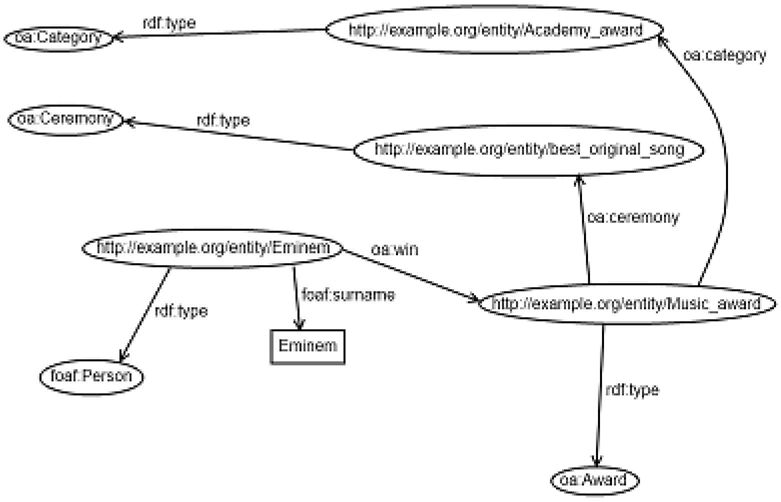
rdf:type
oa:Category
http://example.org/entity/Academy_award
oa:category
oa:Ceremony
rdf:type
http://example.org/entity/best_original_song
oa:ceremony
http://example.org/entity/Eminem
oa:win
foaf:surname
rdf:type
http://example.org/entity/Music_award
Eminem
foaf:Person
rdf:type
oa:Award

Fig. 2. Music example schema

```
ex:eminem rdf:type foaf:Person ;
    foaf:surname "Eminem";
    awards:win ex:Music_award .

ex:Music_award> rdf:type awards:Award ;
    awards:ceremony ex:Academy_award ;
    awards:category ex:best_original_song .

ex:Academy_award rdf:type awards:Ceremony .

ex:best_original_song rdf:type awards:Category .
```

While the ontology has been specifically designed for music, we realised that it could be used for describing awards in other domain (sport, best paper conference, film, etc.), as demonstrated by the following example and its corresponding Turtle code.

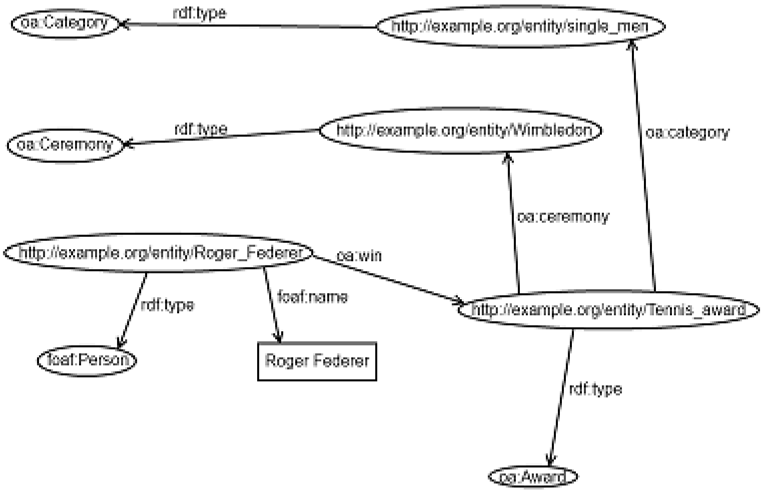
oa:Category
rdf:type
http://example.org/entity/single_men
oa:Ceremony
rdf:type
http://example.org/entity/Wimbledon
oa:category
oa:ceremony
http://example.org/entity/Roger_Federer
oa:win
http://example.org/entity/Tennis_award
rdf:type
foaf:name
foaf:Person
Roger Federer
rdf:type
oa:Award

Fig. 3. Sport example schema

```
ex:Roger_Federer rdf:type foaf:Person ;
    foaf:name "Roger Federer";
    awards:win ex:Tennis_award .

ex:Tennis_award rdf:type awards:Award ;
    awards:ceremony ex:Wimbledon ;
    awards:category ex:single_men .

ex:Wimbledon rdf:type awards:Ceremony .

ex:single_men rdf:type awards:Category .
```

4 Award extraction

While this ontology can be used on top of any data, we also focused on extracting relevant awards information from Wikipedia mapped to this model. On Wikipedia, artists are assigned several categories, some of them containing the word “award”. Our goal here was to transform these categories (as sentences) into the corresponding award information (As structured data), using the previous ontology.

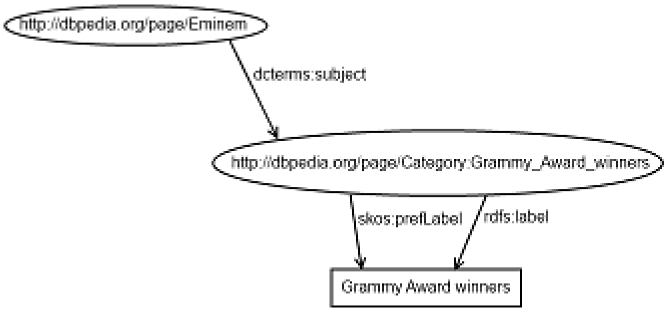
http://dbpedia.org/page/Eminem
dcterms:subject
http://dbpedia.org/page/Category:Grammy_Award_winners
skos:prefLabel
rdfs:label
Grammy Award winners

Fig. 4. Schema representing this link between a person and a string category

The main task was to turn the information embedded in these sentence categories into triples, designed with the award ontology. We used Python, and in particular NLTK (Natural Language ToolKit) as a NLP framework.

NLTK allowed us to extract a piece of a sentence with a method called *chunking*, by creating several regular expressions corresponding to tags associated to each words (noun, adjective, etc.) or words sequence to extract. We split the task in two parts: (1) cremony extraction, and (2) category extraction. Yet, taken alone, the chunking part was not enough sufficient to have a good percentage of correct answer for the extracted ceremonies. To improve the results, we then compared the results with existing ceremonies (defined in Wikipedia), by retrieving those from DBpedia and computing the Levenshtein distance between our results (for each ceremony name retrieved) and the list of ceremonies in Wikipedia/DBpedia, taking the best matching as a result. Finally, to improve the results that were only about 50% using this pipeline, we used Machine Learning, with the Google Prediction API[10], to train the algorithms to differentiate between correct and wrong answers. Our training set as defined as a set of mappings between the NLTK extraction and the correct ceremony name, as follows:

```
NLTK extracted ceremony;dbpedia ceremony;yes or no
```

We were eventually able to reach a score of 69% accuracy for the mappings after using this technique[11]. Here is an example of successful extraction with the process.

```
sentence : ARIA Award winners
extracted ceremony : ARIA Music Awards
```

And another more complete example, including both ceremony and category extraction:

[10] https://developers.google.com/prediction/

[11] The tests have been run on all the corpus of the Wikipedia categories containing the term *award*

```
sentence: worst actress golden raspberry award winners
extracted ceremony: golden raspberry awards
extracted category: worst actress
```

Corresponding to the following representation with the awards ontology:

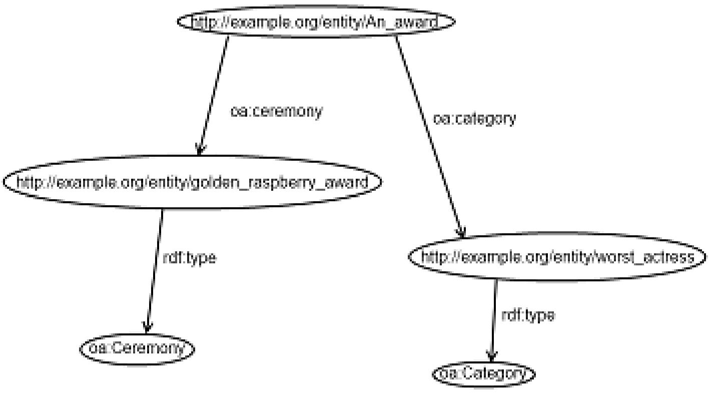
http://example.org/entity/An_award
oa:ceremony
oa:category
http://example.org/entity/golden_raspberry_award
rdf:type
http://example.org/entity/worst_actress
rdf:type
oa:Ceremony
oa:Category

Fig. 5. Schema of the example

```
ex:An_award rdf:type awards:Award ;
    awards:ceremony ex:golden_raspberry_award ;
    awards:category ex:worst_actress .

ex:golden_raspberry_award rdf:type awards:Ceremony .

ex:worst_actress rdf:type awards:Category .
```

It's easy after to know at which artist this extraction can be associated.

5 Conclusion

In this paper, we addressed two main questions regarding events, with a particular focus on the concepts of awards:

- how to represent awards, a particular kind of events, on the Web of Data;
- how to extract award information from semi-structured sources, and to represent it as structured data on the Web

This extraction is currently integrated into seevl, a Chrome extension for music discovery on YouTube. Music fans will consequently be able to identify videos of artists having won particular awards, and to generate related playlists. In the future, we plan to extend the extraction to identify the year related to the awards, and to focus on Wikipedia pages content (and not only category information), to extract this data.

References

1. Martin Doerr. The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, 24(3):75–92, 2003.
2. Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers, 2011.
3. Knud Mller, Livia Predoiu, and Daniel Bachlechner. Portal ontology. Technical report, DERI, 2004.
4. Ansgar Scherp, Thomas Franz, Carsten Saathoff, and Steffen Staab. F–a model of events based on the foundational ontology dolce+DnS ultralight. In Yolanda Gil and Natasha Fridman Noy, editors, *K-CAP*, pages 137–144. ACM, 2009.
5. Ryan Shaw, Raphaël Troncy, and Lynda Hardman. LODE: Linking Open Descriptions of Events. In Asunción Gómez-Pérez, Yong Yu, and Ying Ding, editors, *ASWC*, volume 5926 of *Lecture Notes in Computer Science*, pages 153–167. Springer, 2009.
6. Maria Vargas-Vera and David Celjuska. Ontology-driven Event Recognition on Stories. Technical report, KMI, the Open University, 2003.