A Multilingual/Multimedia Lexicon Model for Ontologies

Paul Buitelaar[1], Michael Sintek[2], and Malte Kiesel[2]

[1] DFKI GmbH, Language Technology, Stuhlsatzenhausweg 3,
66123 Saarbruecken, Germany
paulb@dfki.de
[2] DFKI GmbH, Knowledge Management, Erwin-Schrödinger-Straße,
67608 Kaiserslautern, Germany
{sintek, kiesel}@dfki.de

Abstract. Ontology development is mostly directed at the representation of domain knowledge and much less at the representation of textual or image-based symbols for this knowledge, i.e., the multilingual and multimedia lexicon. To allow for automatic multilingual and multimedia knowledge markup, a richer representation of text and image features is needed. At present, such information is mostly missing or represented only in a very impoverished way. In this paper we propose an RDF/S-based lexicon model, which in itself is an ontology that allows for the integrated representation of domain knowledge and corresponding multilingual and multimedia features.

1 Introduction

Ontologies define the semantics for a *set of objects* in the world using a *set of classes*, each of which may be identified by a particular *symbol* (either linguistic, as image, or otherwise). In this way, ontologies cover all three sides of the "semiotic triangle" that includes *object, referent,* and *symbol*, i.e., an *object* in the world is defined by its *referent* and represented by a *symbol* (Ogden and Richards, 1923 – based on Peirce, de Saussure and others).

Currently, ontology development and the Semantic Web effort in general have been mostly directed at the *referent* side of the triangle, and much less at the *symbol* side. To allow for automatic multilingual and multimedia knowledge markup a richer representation is needed of the linguistic and image-based symbols for the object classes that are defined by the ontology. At present, such information is mostly missing or represented only in a very impoverished way, leaving the semantic information in an ontology without a grounding to the human cognitive and linguistic domain. For instance, according to the collection of ontologies available through OntoSelect[1] (see Buitelaar et al., 2004), currently only about 9% of ontologies represent multilingual terms for classes and/or properties.

Linguistic symbols, i.e., simple words or more complex terms, are represented in a lexicon that provides the meaning of these words or terms, besides a more or less

[1] http://olp.dfki.de/OntoSelect/

Y. Sure and J. Domingue (Eds.): ESWC 2006, LNCS 4011, pp. 502 – 513, 2006.

extensive representation of their linguistic features, e.g., if the word is a noun or a verb, if it is atomic or can be split into multiple words, etc. Similarly, a lexicon of images can be defined that represent which prototypical image, or more precisely, which set of image features corresponds to which ontology class. Here, we will discuss a multilingual/multimedia lexicon model that will allow for the representation of linguistic and image symbols for ontology classes and properties.

2 Ontologies and Multilingual/Multimedia Features

An ontology describes a knowledge model of a particular domain of discourse at a particular point of time and is shared between two or more actors in the domain. As the ontology defines the agreed semantics of the domain, all relevant content will be marked-up with knowledge according to the ontology. The definition of the ontology in turn depends primarily[2] on the content that has already been interpreted. Accordingly, content production and interpretation will drive the adaptation of the ontology infrastructure, and ontology adaptation will drive content interpretation and production.

In order to arrive at such a continuous 'hermeneutic cycle' of content and knowledge production and interpretation, a rich representation of domain knowledge and content features is needed. Here we propose an integrated approach that organizes content and knowledge in several layers:

- *content layer* (outermost layer)
  This layer consists of multilingual (text documents) and multimedia data (images, video and/or mixed image and text documents).
- *features layer* (1st inner layer)
  This layer consists of extracted features for the data in the content layer. For multilingual data, this ranges from comparatively informal feature vectors gathered by use of statistical methods to formalized descriptions of the content of text documents, typically extracted by use of natural language processing and information extraction methods. For multimedia data, this will be mostly limited to informal features as used in color histograms and similar.
- *feature association layer* (2nd inner layer)
  This layer consists of ontology-based representations of the multilingual and multimedia features also occurring in the features layer. While in the *features layer* features are associated with multilingual and multimedia data, in the *feature association layer* the features are associated with ontology classes and relations.
- *ontology layer* (central layer)
  This layer consists of ontology classes and relations, with which the data in the content layer is to be interpreted (i.e., annotated) by use of the extracted and represented features in the *features layer* and the *feature association layer*.

[2] Aside from more generic knowledge of the physical world, time, space, etc. that will be inherited from an upper-level ontology.

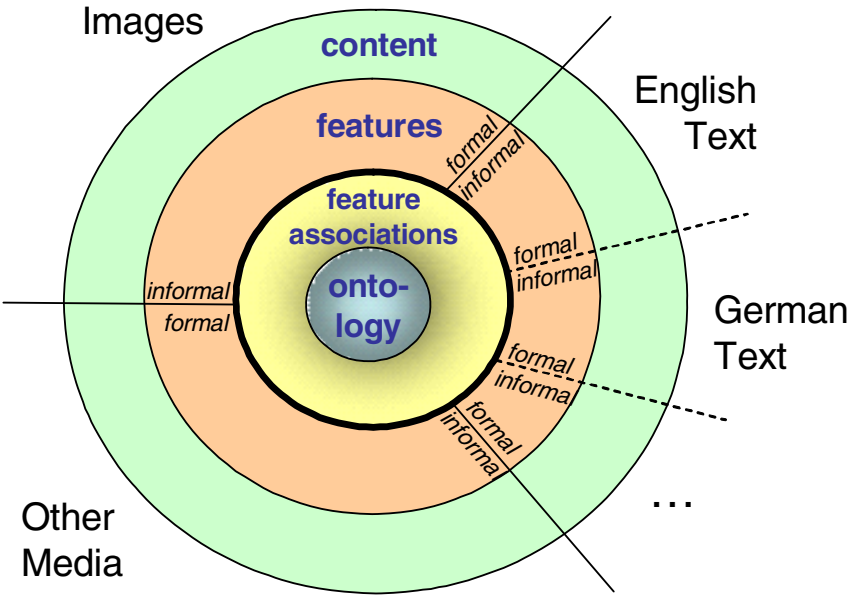

Fig. 1. Interacting Layers in Feature Extraction and Representation

3 Towards an Ontology-Based Representation of Multilingual and Multimedia Features

In the following, we describe how to represent multilingual and multimedia features in ontologies and how to link them to ontology concepts.

3.1 Representation of Multilingual and Multimedia Features

Multilingual features consist of a list of term variants - for each language covered by the ontology - with lexical and context information for each term:

- *language-ID*: ISO-based unique identifier for the language of each term
- *part-of-speech*: (possibly ISO-based) representation of the part of speech of the head of the term
- *morphological decomposition*: representation of the morphological structure (segments, head, modifiers) of a term
- *syntactic decomposition*: representation of the syntactic structure (segments, head, modifiers) of a term
- *statistical and/or grammatical context model*: representation of the linguistic context of a term in the form of N-grams, grammar rules or otherwise

Multimedia features will be represented by MPEG-7 descriptors (see also Petridis et al., 2004) for properties such as:

- *color*: color space, structure, layout; dominant color, scalable color
- *texture*: homogeneous texture, texture browsing, edge histogram
- *shape*: contour-based, region-based, 3-D, multiple-views

3.2 Annotating Ontology Classes with Multilingual and Multimedia Features

To represent terminology in different languages as well as multimedia features, we created an RDF/S-based domain knowledge representation introducing meta-class `ClassWithFeats` and meta-property `PropertyWithFeats`, as shown in Figure 2. Using meta-classes and meta-properties allows us to connect content features to classes and properties directly. In ontology tools such as Protégé (Noy et al., 2001), using `ClassWithFeats` as meta-class for a domain class results in additional widgets getting displayed along with the standard class widgets such as Name and Documentation. In these new widgets, the features of the corresponding class or property can be entered, populating the `feat:lingFeat` and `feat:imgFeat` properties for each class.

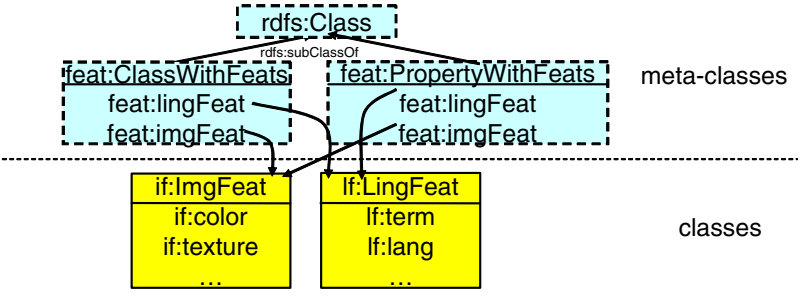

Fig. 2. ClassWithFeats and PropertyWithFeats

For instances, we attached the `feat:lingFeat` property to the root class of the domain ontology. This way every instance of the knowledge base can get annotated with linguistic information, e.g., allowing representation of language-dependent names. The same can be done with the `feat:imgFeat` property.

The integrated ontology-based feature representation we propose is based on ongoing work in the context of the SmartWeb[3] project on mobile Semantic Web access for intelligent information services in the soccer domain. The proposed feature representation is currently used in the SmartWeb ontology on sports events and related issues (see also section 5).

Figure 3 shows the ontology with example (domain) classes and associated linguistic and image features: the ontology contains the class `o:FootballPlayer` with subclasses `o:Defender` and `o:Midfielder`. All these classes are instances of the meta-class `feat:ClassWithFeats` which allows them to use the feature-association properties `feat:lingFeat` and `feat:imgFeat`.

[3] http://www.smartweb-project.de/

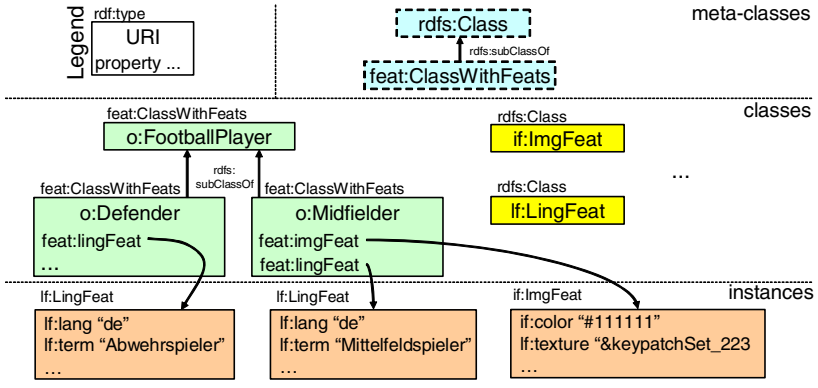

Fig. 3. Ontology and Examples (simplified) – *Defender, Midfielder*

Figure 4 depicts the part of our ontology in detail that deals with the representation of linguistic features, which is mainly the morphosyntactic decomposition of phrases and word forms down to stems, roots, morphemes, affixes etc. Apart from having linguistic properties like gender, number, part of speech, case, etc., word forms have the property `semantics` which is a back link into the ontology allowing semantics to be assigned to them.

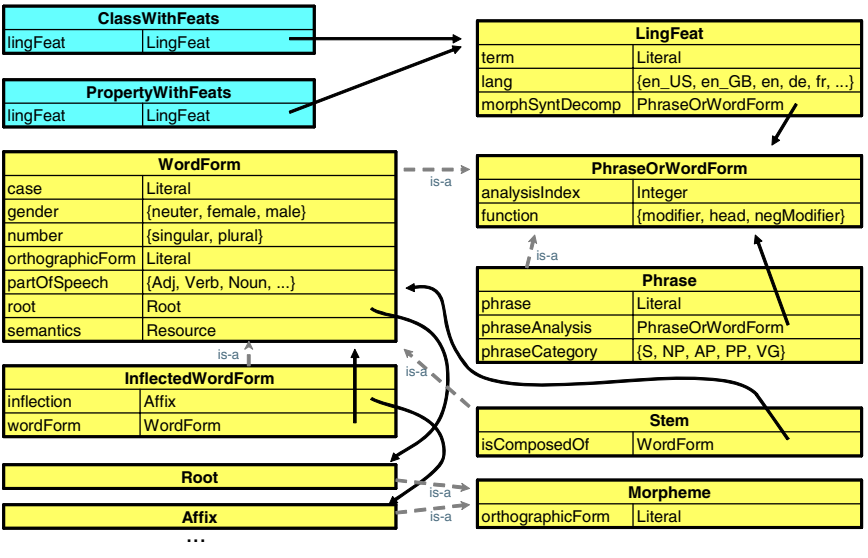

Fig. 4. Linguistic Features in Detail

Figure 5 shows a sample application of this part of the ontology, the decomposition of the German term "Fußballspielers" (= "of the football player"): `inst1` indicates that is an inflected word form (where the inflection is for forming the genitive) with

stem "Fußballspieler" (`inst2`, "footballplayer"), which can be decomposed into two stems, "Fußball" (`inst3`, "football") and "Spieler" (`inst8`, "player"); this is recursively continued for "Fußball" which is composed of the stems "Fuß" and "Ball" (`inst5` and `inst7`, "foot" and "ball").

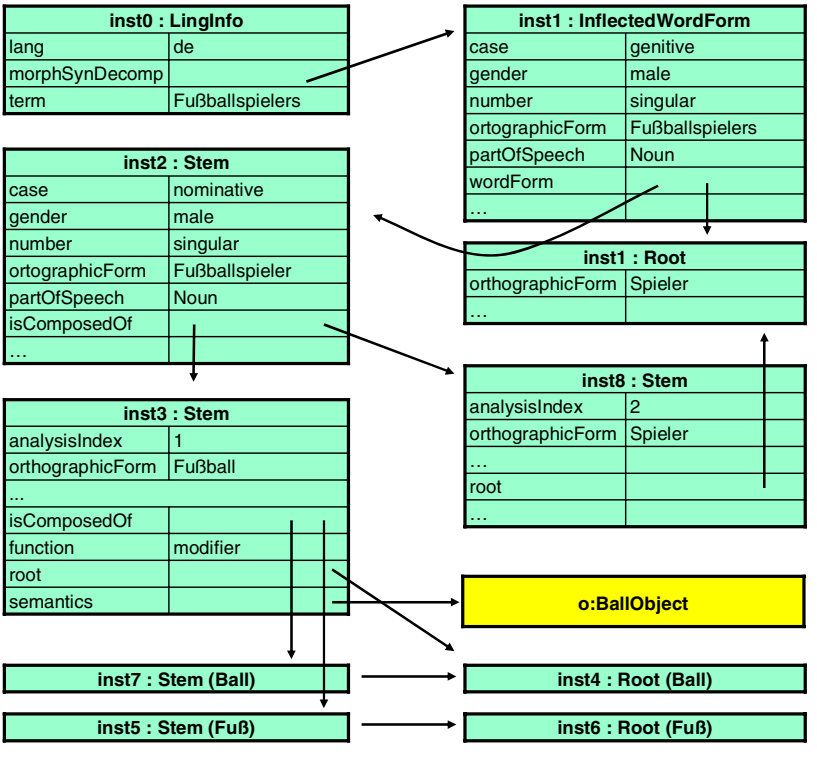

Fig. 5. Morphosyntatic Decomposition of "Fußballspielers"

4 Comparison with Related Work

The multilingual/multimedia lexicon model we propose has some overlap with related proposals, of which we discuss the most prominent ones here:

- SKOS: Simple Knowledge Organization System
- OntoWordNet
- LMF: Lexical Markup Framework

Of these, SKOS originates out of the W3C working group on "Best Practices for the Semantic Web"[4], whereas LMF is a working draft of the ISO working group on Language Resources Management TC37/SC4[5] (Francopoulo, 2006).

[4] http://www.w3.org/2001/sw/BestPractices/
[5] http://www.tc37sc4.org

4.1 SKOS - Simple Knowledge Organization System

Although there is some overlap with SKOS[6] (Miles and Brickley, 2005a, 2005b), the proposed representation is richer as it will include not only multilingual terms for classes (and properties) but also multimedia features and context models.

However, more specifically there is also a technical and conceptual reason why SKOS does not fulfill the needs of our scenario[7]: SKOS uses sub-properties of `rdfs:label` (`skos:prefLabel`, `skos:altLabel`) together with `xml:lang` to attach multilingual terms to concepts.

Furthermore, the RDFS specification[8] (Brickley and Guha, 2004; Hayes, 2004) defines the range of `rdfs:label` to be `rdfs:Literal`. From the definition of `rds:subPropertyOf` follows that the range of `skos:prefLabel` and `skos:altLabel` is also `rdfs:Literal` (or a specialization of `rdfs:Literal`). This is not sufficient in our scenario since we want to attach more information as linguistic information to classes than simple multilingual strings. This led to our decision to use the meta-class `ClassWithFeats`, which allows us to attach complex information to classes with the properties `lingFeat` and `imgFeat` (in the future, more properties will be defined for other media types like audio and video).

The conceptual problem we see with SKOS for the use in our scenario is that it mixes linguistic and semantic knowledge. SKOS uses `skos:broader` and `skos:narrower` to express "semantic" relations without clearly stating the semantics of these relations intentionally, and defines the sub-properties `skos:broaderGeneric` and `skos:narrowerGeneric` to have class subsumption semantics (i.e., they inherit the `rdfs:subClassOf` semantics from RDFS). We clearly keep the linguistic and semantic, ontology-based knowledge representations apart: the ontology is represented using the semantic relations defined in RDFS or OWL-Full[9] (McGuinnes and van Harmelen, 2004), and attach linguistic knowledge to the classes (and properties).

We further propose to integrate image-related features in this representation, which is beyond the scope of SKOS. Note that SKOS uses `foaf:depiction`, `skos:prefSymbol`, and `skos:altSymbol` to attach images to concepts, but not complex feature descriptions.

4.2 Wordnets and OntoWordNet

Our approach in effect integrates a domain-specific multilingual wordnet into the ontology, although also the wordnet model does not distinguish clearly between linguistic and semantic information - see e.g. (Miller et al., 1995) on WordNet and (Vossen, 1998) on EuroWordNet.

---

[6] http://www.w3.org/TR/swbp-skos-core-guide/

[7] In fact, our argumentation applies to all approaches based on `rdfs:label` and `xml:lang` to attach multilingual labels to classes and relations.

[8] http://www.w3.org/TR/rdf-schema/

[9] OWL Lite and OLW DL do not support meta-classes and meta-properties (see http://www.w3.org/TR/owl-features/)

Alternative lexicon models that are more similar to our approach include (Bateman et al., 1995; Alexa et al., 2002), but these concentrate on the definition of a top ontology for lexicons instead of text/image features for domain ontology classes and properties as in our case. This is also the main difference with the proposed OntoWordNet model (Gangemi et al., 2003), which aims at merging the foundational ontology DOLCE (Gangemi et al., 2002) with WordNet to provide the latter with a formal semantics.

4.3 LMF – Lexical Markup Framework

Closest to our goals is the LMF or Lexical Markup Framework by the ISO working group on Language Resources Management TC37/SC4. "The goals of LMF are to provide a common model for the creation and use of very large scale lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of large numbers of different individual electronic resources to form large global electronic resources. … The ultimate goal of LMF is to create a modular structure that will enable true content interoperability across all aspects of lexical resources."

The main difference with LMF and the lexicon model proposed here is the strict division of linguistic and semantic knowledge. In LMF these are integrated in the same model by way of a lexical semantics slot, whereas in our model all lexical semantics is to be found in the domain ontology - that is outside of the lexicon model per se. On top of this, our model allows also for the representation of non-linguistic, i.e. multimedia features.

Nevertheless, the aims and structure of LMF and our model are sufficiently similar to investigate ways of merging the two proposals. We envision this as a potential enrichment on both sides, as our model has a more principled approach to knowledge representation that builds directly on current standards in this area (i.e. RDFS), whereas the LMF model has a strong background in the representation of linguistic knowledge.

5 Applications

The integrated LingInfo approach allows for cross-lingual, cross-media feature extraction, representation and employment as follows:

- *text2image - cross-lingual acquisition of German content features by use of represented English content features*
  i.e., if we know which terms express a class in English then we can build a classifier for the classification of images that occur in the context of English terms for this class
- *image2text - cross-media acquisition of German content features by use of represented multimedia features*
  i.e., if we know which images represent instances for a specific class then we can extract German terms for this class from surrounding German text
- *text2text - cross-media acquisition of multimedia content features by use of represented English content features*

i.e., if we know which terms express a class in English and the context features (i.e. words) for these terms and possible translations into German then we can build a cross-lingual classifier for recognition of unseen German terms for this class

- *text2class, image2class - data-driven adaptation of domain knowledge representation for a class by use of represented English terminology*
  i.e., if we know which terms express a class in English and the context words for these terms then we can detect a change in the semantic model for this class by monitoring any change in the context words - similar with image feature models

5.1 Application of LingInfo in SmartWeb

LingInfo is developed and used within the SmartWeb project, which aims at the development of a complex multi-modal question answering and dialog system that derives answers from unstructured resources such as the Web, from automatically acquired knowledge bases and from web services.

A central component is SWIntO, the SmartWeb Integrated Ontology (Oberle et al. in prep.), which consists of three layers: the upper model DOLCE (Gangemi et al., 2002), the domain-independent model SUMO (Niles and Pease, 2001) the SportEvents ontology, focused mainly on soccer, and further task ontologies. The SportEvents ontology contains about 400 direct classes, all of which are provided with linguistic information as described above.

Enriching the ontology with linguistic information is an incremental process, by which some information can be derived semi-automatically from annotated corpora. In this way, lexicons (and grammars) of available tools are in effect tuned to the soccer domain and become fully integrated with the SmartWeb ontology. Alternatively, if such resources cannot be integrated into LingInfo (e.g. due to copyright problems), pointers may be used to refer to external resources.

Multimedia information is not yet being added to the ontology on a larger scale, but also here a semi-automatic approach will be explored that exploits automatically annotated image collections - where the annotation is performed on the basis of the textual context of the images (Buitelaar et al. 2006).

5.2 LingInfo in Information Extraction from Text

In the SmartWeb project, the LingInfo model is interfaced with the information extraction (IE) system SProUT (Drozdzynski et al., 2004). Based on the information encoded in LingInfo, we automatically extract gazetteer entries for named entities, with back-references to the ontology. For terms associated with concepts, we recompile the relevant parts of the ontology, including LingInfo, into a type hierarchy used in the IE system. Thus, LingInfo information can be used to consistently identify and mark up (inflected) occurrences of domain-relevant terms. The following example may illustrate this. It displays an excerpt of the SWIntO ontology that has been compiled into a type hierarchy defined in TDL[10], the representation language used by SProUT:

[10] Type Description Language – see (Krieger and Schäfer 1994) for details.

```
PlayerAction :< SportMatchAction.

SingleFootballPlayerAction :< PlayerAction.

FootballTeamAction :< PlayerAction.

GoalKeeperAction :< SingleFootballPlayerAction.

AnyPlayerAction :< SingleFootballPlayerAction.
```

Properties associated with these concepts are translated to TDL *attributes* of the corresponding *types*, e.g. the property `inMatch` of the SWIntO class `SportMatchAction` translates to the TDL attribute `INMATCH` that is inherited by all subtypes of the TDL type `SportMatchAction`. The SWIntO property `CommittedBy` that is defined for the SWIntO class `SingleFootballPlayerAction` translates to a corresponding TDL attribute `COMMITTEDBY` of the TDL type `SingleFootballPlayerAction`, and is again inherited by all its subtypes:

```
SportMatchAction := swinto_out & [INMATCH Football].

SingleFootballPlayerAction := swinto_out & [COMMITTEDBY
FootballPlayer].
```

Multilingual (e.g. German) terms that are encoded as LingInfo instances are compiled into TDL lexical types:

```
"Teamaktion" :< FootballTeamAction.

"Spieleraktion" :< PlayerAction.

"Torwartaktion" :< GoalkeeperAction.

"Gesperrt" :< Banned.
```

SProUT extraction patterns can thus be triggered by lexical types, and define output structures that correspond directly to the classes and properties of the SWIntO ontology. For instance, the extraction rule below matches an extraction pattern for the SWIntO (SportEvents) class `BanEvent` with attributes `CommittedBy` and `InMatch` that is triggered for instance by the German LingInfo term "gesperrt". Example sentences from the SmartWeb development corpus[11] to which this rule applies are as follows:

"… ist Petrow für die Partie gegen Schweden gesperrt." ("… has Petrow been banned for the match against Sweden")

"… ist David Trezeguet von der FIFA für zwei Spiele gesperrt worden." ("… has David Tezeguet been banned by FIFA for two matches")

[11] See also http://www.dfki.de/sw-lt/olp2_dataset/

```
banned_player :>

@seek(player) & [IMPERSONATEDBY #player, INMATCHTEAM #team1]

(@seek(weekday_only) & [DOFW #dofw])? (token{0,2}
 @seek(soccer_institutions))?  token{0,3}
 @seek(game_teams) & [INTOURNAMENT #tour, TEAM2 #team2] morph & [STEM banned, SURFACE #event])

-> playeraction &
   [SPORTACTIONTYPE #event,
    COMMITTEDBY footballplayer &
            [IMPERSONATEDBY #player],
    INMATCH match &
            [INTOURNAMENT #tour, MATCHTYPE #match, TEAM1 #team1, TEAM2 #team2]].
```

Fig. 6. SProUT Extraction Rule for the SWIntO Class `BanEvent`

6 Conclusions and Future Work

In this paper we proposed a model for the representation of multilingual and multimedia content features in ontologies, which will allow for more efficient automatic processing of textual and image data in knowledge markup, ontology learning and other applications such as dialog processing, summarization, machine translation, etc.

The model we propose clearly separates domain knowledge on sets of objects from linguistic- and image-related knowledge on terms and images used for referring to such objects. In this way, our proposal extends traditional knowledge representation models used in ontology definition as well as current models used in defining computational lexicons (i.e. Wordnets) and thesauri (i.e. SKOS).

In future work we also intend to expand the model towards the representation of multilingual and multimedia content features for instances. In this way, the knowledge base for a given ontology will be able to represent the linguistic and/or image context for extracted facts.

Acknowledgements

This research has been supported in part by the SmartWeb project, which is funded by the German Ministry of Education and Research under grant 01 IMD01 A.

References

M. Alexa, B. Kreissig, M. Liepert, K. Reichenberger, L. Rostek, K. Rautmann, W. Scholze-Stubenrecht, S. Stoye *The Duden Ontology: an Integrated Representation of Lexical and Ontological Information* In: Proc. of the OntoLex Workshop at LREC, Spain, May 2002.

J. A. Bateman, R. Henschel and F. Rinaldi *Generalized Upper Model 2.0: documentation* Report of GMD/Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, Germany, 1995.

D. Brickley, R.V. Guha (eds.) RDF Vocabulary Description Language 1.0: RDF Schema. World Wide Web Consortium, 2004.

P. Buitelaar, Th. Eigner, Th. Declerck *OntoSelect: A Dynamic Ontology Library with Support for Ontology Selection* In: Proc. of the Demo Session at the International Semantic Web Conference, Hiroshima, Japan, Nov. 2004.

P. Buitelaar, P. Cimiano, S. Racioppa and M. Siegel *Ontology-based Information Extraction with SOBA* In: Proc. of the International Conference on Language Resources and Evaluation (LREC), 2006.

W. Drozdzynski, H.-U. Krieger, J. Piskorski, U. Schäfer, F. Xu *Shallow Processing with Unification and Typed Feature Structures - Foundations and Applications.* In Künstliche Intelligenz, 1/2004.

G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, C. Soria *Lexical Markup Framework (LMF)* In: Proc. of the International Conference on Language Resources and Evaluation (LREC), 2006.

A. Gangemi, Guarino, N., Masolo, C., Oltramari, A. and L. Schneider. 2002. *Sweetening Ontologies with DOLCE.* In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW), Siguenza, Spain, pp. 166-181.

A. Gangemi, Navigli R, Velardi P *The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet.* Meersman R, et al. (eds.), Proceedings of ODBASE03 Conference, Springer, 2003.

P. Hayes (ed.) *RDF Semantics.* World Wide Web Consortium, 2004.

H.-U. Krieger and U. Schafer *TDL---a type description language for constraint-based grammars* In Proceedings of the 15th International Conference on Computational Linguistics (COLING), pp. 893-899, 1994.

D.L. McGuinness, F. van Harmelen (eds.) *OWL Web Ontology Language Overview.* W3C Recommendation 10 February 2004.

A. Miles, D. Brickley (ed.) *SKOS Core Vocabulary Specification.* W3C Working Draft 10 May 2005a.

A. Miles, D. Brickley (eds.) *SKOS Core Guide.* W3C Working Draft 10 May 2005b.

G. A. Miller *WORDNET: A Lexical Database for English*. Communications of ACM (11): 39-41, 1995.

I. Niles and Pease, A. *Towards a standard upper ontology*. In: FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems, ACM Press (2001)

N. F. Noy, M. Sintek, S. Decker, M. Crubezy, R. W. Fergerson, & M. A. Musen. *Creating Semantic Web Contents with Protege-2000*. IEEE Intelligent Systems 16(2):60-71, 2001.

D. Oberle, A. Ankolekar, P. Hitzler, P. Cimiano, C. Schmidt, M. Weiten, B. Loos, R. Porzel,H.-P. Zorn,M. Micelli, M. Sintek,M. Kiesel, B. Mougouie, S. Vembu, S. Baumann, M. Romanelli, P. Buitelaar, R. Engel, D. Sonntag, N. Reithinger, F. Burkhardt, J. Zhou *DOLCE ergo SUMO: On Foundational and Domain Models in SWIntO (SmartWeb Integrated Ontology),* in preparation.

Ch. K. Ogden and I. A. Richards *The meaning of meaning - A study of the influence of language upon thought and of the science of symbolism.* London: Kegan Paul, Trench, Trubner & Co., 1923.

K. Petridis, I. Kompatsiaris, M. G. Strintzis, S. Bloehdorn, S. Handschuh, S. Staab and N. Simou *Knowledge Representation for Semantic Multimedia Content Analysis and Reasoning* In: Proc. of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, Royal Statistical Society, London, 25-26 Nov. 2004.

Vossen P. (ed). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht, 1998