ADDING SEMANTIC METADATA TO AUDIO-VIDEO MATERIAL BY AUTOMATIC ANALYSIS OF COMPLEMENTARY SOURCES

T. Declerck*, P. Buitelaar*, M. Alcantara*, M. Labský†, V. Svátek†

*Deutsches Forschungszentrum für Künstliche Intelligenz
Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany
E-mail: declerck@dfki.de

†University of Economics, Prague
Nam.W.Churchilla 4, 13067 Praha 3, CZECH REPUBLIC
Email: Svatek@vse.cz

Keywords: Human Language Technology, Multimedia Semantics, Semantic Web.

Abstract

We present in this paper actual work on adding semantic metadata to multimedia material, on the base of the results of the automatic analysis applied to associated language material, being speech transcripts or various types of textual documents related to video/image material

1 Introduction

The paper describes some approaches dealing with the use of natural language processing (NLP) techniques for semantically indexing or annotating multimedia (MM) material. Those approaches are currently being investigated in the European Network of Excellence "K-Space" (see http://www.k-space.eu/). We will describe the kind of annotation that can be delivered by typical modules of text analysis systems when applied to textual material accompanying multimedia material (we will use the term of *complementary sources* for those documents).

More substantially, the paper addresses the question why using NLP for indexing/annotating multimedia material. For the time being, the automatic analysis of video/image material is mostly resulting in so-called "low-level" content features (colour, texture, shape etc.). Comparing to the way humans perceive and access MM content, we say that there is a "semantic gap" in the field of automatic content detection (and indexing) in video/image processing. The integration of semantics encoded in associated speech and/or text (superposed text, caption, subtitles, continuous text etc.) or other available modalities (gestures etc.) might offer some help in reducing this semantic gap.

Various NLP techniques can be used for analyzing text associated to image/video material. So for example robust NLP techniques can be used for indexing MM material with key words or named entities. Those robust techniques include Automatic Speech Recognition (ASR), the OCR processing of text in images or the processing of short text sequences in subtitles of videos and captions of images. In those cases, single annotation structures for indexing MM material are generated. For more ambitious tasks, like the event annotation of MM material, the use of more advanced NLP technologies is required. Some systems here combine language technology and domain models, for example in the form of ontologies, for supporting the extraction of relevant entities, relations and events from various types of textual documents in various languages. Here the text analysis is able to generate complex annotation structures, describing relevant (sequences of) events.

The challenge today is to go for a automated precise alignment of high-level semantic features gained from the analysis of complementary sources and the low/mid-level features extracted by audio-video analysis, supporting at the end a real cross-media search through large archives or through the web. Some running projects, like the Integrated Project (IP) aceMedia, are investigating the use of ontologies to level low-level features to a higher level of abstraction. "K-Space", the IP "MESH", or the German project "SmartWeb" are investigating the development of an integrated knowledge structure for text, speech, audio and video descriptions. We therefore also describe in this paper also the building of an integrated data set that will be supporting the automatic alignment of features resulting from text, speech and image analysis.

This work reported in this paper is also very relevant for the other direction of investigation: using image analysis for improving textual information extraction from the web, as done in the Czech project "Rainbow", on which we report briefly at the end of the paper.

2 Multimedia Semantics

The topic of Multimedia Semantics has gained a lot of interest in recent years, and large funding agencies issued calls for R&D proposals on those topics. So for example the 4th call of the 6th Framework of the European Commission was dedicated to the merging of results gained from R&D projects on knowledge representation on the one hand and projects dedicated to cross-media content on the other hand. The goal

being in making the (semantic) descriptions of multimedia content re-usable on the base of a higher interoperability of media resources, which has been so far described mainly at the level of XML syntax, as can be seen with the MPEG-7 standard for encoding and describing multimedia content.

2.1 Ontologies for Low-Level Features

In the line of the recent developments in the fields of Semantic Web technologies, one approach consists in looking at ways for encoding so-called low-level features, as they can be extracted from audio-video material, into a high-level features organization as one can typically find in (domain) ontologies. The EC co-funded project aceMedia ([16]) is offering a very good example of such an approach. In this project, ontologies, which are typically describing knowledge as expressed in words, are extended in order to include the low-level visual features resulting from state-of-the-art audio-video analysis systems. For the description of low-level features, the project uses as its background the MPEG-7 standard, and proposes links from the MPEG-7 descriptors to high-level (domain) ontologies (see [1]).

2.2 An integrated Approach: The K-Space Network of Excellence

The European Network of Excellence "K-Space" (Knowledge Space of Shared Technology and Integrative Research to Bridge the Semantic Gap), which started in 2006, is dealing with semantic inferences for semi-automatic annotation and retrieval of multimedia content. The aim is to narrow the gap between content descriptors that can be computed automatically by current machines and algorithms, and the richness and subjectivity of semantics in high-level human interpretations of audiovisual media: the so-called *Semantic Gap*..

The project deals with a real integration of knowledge structures in ontologies and low-level descriptors for audio-video content, taking also into account knowledge that can be extracted from sources that are complementary to the audio/video stream, mainly speech transcripts and text surrounding images or textual metadata describing a video or images. The integration work takes place at 2 levels: the level of knowledge representation, where features associated with various modalities (image, text/speech transcripts, audio) should be interrelated within conceptual classes in ontolologies (from domain-specific to general purpose ontologies), and the level of processing, where high-level semantic features should be integrated for guiding (and so possibly improving) the automatic analysis of audio-video material and the corresponding extraction of semantic features.

An interesting project with respect to K-Space is MESH ([24]), which to a certain extent is building an application scenario (in the News domain) on the base of the kind of multimedia and cross-media knowledge structures discussed and proposed by aceMedia and K-Space, addressing also the issue of harmonization, syndication, personalization and distribution of semantically annotated multimedia and multimodal information material.

The multimedia ontology infrastructure of K-Space is containing qualitative attributes of the semantic objects that can be detected in the *multimedia material*, e.g. colour homogeneity, in the *multimedia processing methods*, e.g. colour clustering, and in the *numerical data or low-level features*, e.g. colour models. The ontology infrastructure will also contain the representation of the top-level structure of multimedia documents in order to facilitate a full-scale annotation of multimedia documents.

Also, a prototype knowledge base will be designed to enable automatic object recognition in images and video sequences. Prototype instances will be assigned to classes and properties of the domain specific ontologies, containing low level features required for object identification.

Partners of K-Space dealing with textual analysis will integrate into this ontology infrastructure the typical features for text analysis, also proposing ontology classes at a higher-level, that supports the modelling of interrelated cross-media features (multimedia and text). We are basing our work on the proposal made by [2].

2.3 Development of a Multimedia Data Set

In the context of the K-Space project we are developing a multimedia data set that will allow for cross-media knowledge acquisition experiments. The multimedia data set is based on a previously compiled data set (the "SmartWeb Data Set" – see also below)[1] in the football (soccer) domain that consists of web-based textual, tabular, and image data on the 2002 and 2006 world cup. To integrate this data set with corresponding A/V data alignment algorithms will be developed, making use of linguistic analysis (for the textual data) image analysis (for the A/V and image data) and semantic analysis (for all types of data) according to a domain ontology. The resulting aligned multimedia data set will then be further exploited in cross-media knowledge acquisition experiments, including multimedia ontology population (acquisition of textual and image instances), ontology enrichment (acquisition of linguistic and image features for ontology classes/properties) and ontology learning (extension of the domain ontology with new classes/properties).
The SmartWeb data set consists of:
An ontology on football that is integrated with foundational (DOLCE), general (SUMO) and task-specific (discourse, navigation) ontologies

- A corpus of semi-structured (tabular) and textual match reports (German and English documents) that are derived from freely available web sources. The bilingual documents are not translations, but are aligned on the

[1] See http://www.dfki.de/sw-lt/olp2_dataset/. This Data Set was developed within the SmartWeb project. See http://www.smartweb-projekt.de/ for more details

level of a particular match, i.e. they are about the same match.

- A knowledge base of events and entities in the world cup domain that have been automatically extracted from the German documents

The SmartWeb Integrated Ontology (SWIntO, see [14]) integrates a selection of domain- and task-specific ontologies with general/foundational ontologies, which results in a relatively deep hierarchical structure with foundational DOLCE classes on top (e.g. Attribute), followed by general SUMO classes (e.g. SocialRole) and finally classes in the soccer domain (e.g. FootballPlayer).

Additionally, to enable easy employment of the ontology in text-based knowledge processing, many of the SWIntO classes carry information on German and English terms and synonyms. To allow for a direct connection of this linguistic information for terms with corresponding classes and properties in the domain ontology, [2] developed a lexicon model (LingInfo) that enables the definition of LingInfo instances (each of which represents a term) for each class or property. The LingInfo model is represented by use of a meta-class `(ClassWithLingInfo)` and meta-property `(PropertyWithLingInfo)`, which allow for the representation of LingInfo instances with each class, where each LingInfo instance represents the linguistic features (`feat:lingInfo`) of a term for a particular class.

Figure 1 shows an overview of the model with example domain ontology classes and associated LingInfo instances. The domain ontology consists of the class `o:FootballPlayer` with subclasses `o:Defender` and `o:Midfielder`, each of which are instances of the meta-class `feat:ClassWithLingInfo` with the property `feat:lingInfo`.

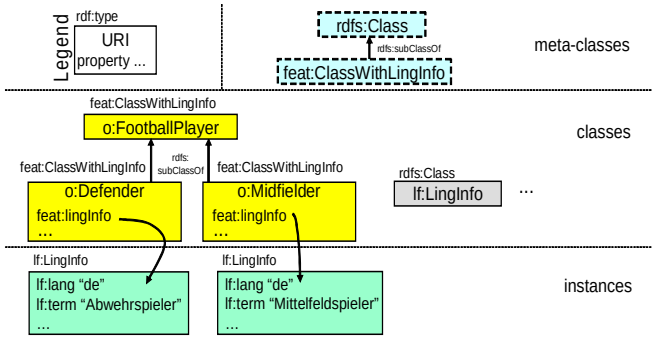

Figure 1: LingInfo model with example domain ontology classes and LingInfo instances (simplified)

The SmartWeb corpus is organized around semi-structured data on each of the qualifying matches for the Soccer World Cup 2006. Semi-structured data consists of a tabular report on the match, to which textual match reports have been automatically assigned. The SmartWeb corpus -- semi-structured, textual match reports and captions in German and English some captions also in French and Spanish -- is available as a set of XML files, with textual match reports and image captions linked to a corresponding semi-structured match report by way of so-called cross-ref XML files (see [15] for details).

The SmartWeb knowledge base has been automatically derived from the German corpus on the basis of named-entity recognition and semantic annotation of events. The SmartWeb knowledge base is available as a set of RDF files, with one RDF file for each match that is linked to corresponding semi-structured and textual match reports by way of the cross-ref files.

In the context of K-Space, the SmartWeb data set is taken as a starting point for developing up a multimedia data set in the football domain. Current work involves the integration of 2006 and historic videos of the 2002 world cup tournament matches, for which event recognition (goals) can be provided and encoded in MPEG7, with textual, tabular and image data on the same matches, as available in the SmartWeb data set.

2.4 Use of Textual Information and Knowledge Bases for Semantic Feature Extraction from Audio Signal

In K-Space some work is dedicated to the extension of state-of-the-art processing and analysis algorithms to handle high-level, conceptual representations of knowledge embedded in audio content based on reference ontologies and semantically annotated associated text, including speech transcripts, when the quality of the transcripts allows it.

The speech transcripts are often encoded using the text annotation tool of the Linguistic Description Scheme (LDS) of MPEG-7. An example of such a (manual) annotation related to a video sequence is given just below, whereas we take this example from a TRECVid data set (year 2005):

```
<VideoSegment id="shot1_13">
  <MediaTime>
  <MediaTimePoint>T00:01:40:11008F30000</MediaTimePoi
  nt>
    <MediaDuration>PT10S26326N30000F</MediaDuration>
  </MediaTime>
  <TextAnnotation confidence="0.500000">
    <FreeTextAnnotation>
      TRACKS STOPPED ROLLING NOSE AND FORMALLY
FILED A HIGHWAY WITH EIGHT DAIL NEW YORK
NEWSPAPERS WHERE THE VOID OF NEWSPAPERS THE
VOID OF CUSTOMERS
    </FreeTextAnnotation>
  </TextAnnotation>
</VideoSegment>
```

Interesting to note in the given example, is that the media time is also given, so that this can be used as a way to look for alignment of the low-level features and the high-level features that can be extracted from the text.

Linguistic analysis of speech transcriptions can help us to understand what videos show, but since ASR systems do not deliver full reliable transcripts, it is necessary to implement methods for getting information from them in spite of their inaccuracies. The highest results (around 90%) of correct words in ASR can be only expected in the best -and most artificial- conditions. ASR quality depends on adaptation processes for speaker models, acoustic environment conditions and speech spontaneity, and none of them can be predicted when working with non-restricted input data.

Another problem is related to the task of applying linguistic annotation to speech transcripts. We have processed the ASR output by means of a POS tagger ([10]) and a chunker ([11]) so that every word has information about its POS and lemma, and some basic phrasal information is given. However, current taggers have been trained on written language texts and are not able to deal with many spontaneous spoken phenomena. To take an example, discourse markers and multi-word units are not taken into account. Besides, there is not a standard annotation for spoken language yet. For solving partially this problem, we will have to implement repair strategies.

With no punctuation, prosodic tagging nor accurate ASR, a successful chunking is not to be expected. However, we use it for the detection of linguistic fragments because groups of words can reduce the global amount of ambiguities present after the tagging of the transcripts.

Once the ASR texts were processed, we analyzed groups with nominal and verbal core (noun and verb phrases respectively). They are considered to be the most relevant word groups for substantive information and, therefore, the easiest way for interpreting shot transcripts.

Spoken language is more redundant than the written one. Repetitions of terms within a shot prove their relevance even when the ASR texts include errors and lack of structure. This method is inspired by previous experiments where repetitions of different fragments were compared to automatically determine segment boundaries ([12]). On the base of the detection of repetitions, we might be able to guess with more evidence the content of the video/image.

Detecting verbal constructions in the transcripts can give us hints about the most basic linguistic structure of a transcript even when ASR output is incomplete and delivered with no structure at all. With this aim, we distinguish between the three event types proposed in SESCO ([9]): states (which are relations between entities and properties/locations), processes (transitions from one state into another) and actions (processes carried out by agents). States are crucial for linking transcriptions and shot contents. Locative states usually relate entities to the places where they appear in images. These references are especially valuable when they are specific (e.g. “the Woodbury Gardens”), but more general locations can also be meaningful (e.g. “California”). Attributive states help us to determine how the entities appear in the shot and their special characteristics (if any). Some states even make deictic references and describe images explicitly. A list of constructions is necessary to find cases such as “This is…”, “Here is…”, “There we see…”, etc. , which seem to be frequent in video corpora ([13]).

Processes and actions have a more complex structure, but they can be used in a similar way. World knowledge ontologies (WorldNet and domain specific ontologies) help us to locate the image in those cases in which processes and actions are clearly linked to places (e.g. “shopping” > “shop”, “teaching” > “classroom”, etc.). There are also deictic references (e.g. “let’s move to…”), but far less frequent than in states. Cases of processes and actions not easily related to their shots seem to be much more common than those of states.

To summarize: Quality of current ASR systems does not allow complex linguistic analysis, but they are already accurate enough to give hints about the video contents. We have used very shallow analysis till now, but more complex approaches have to be added in order to improve the results:

- First, we need a database of lemmas related to its event structures. Transcriptions are too fragmentary to try a traditional syntactic analysis, but events give useful basic information about linguistic structures. Since it is a time consuming task, the database should be focused on the most frequent verbs found in multimedia corpora.
- Secondly, these structures have to be linked to world knowledge through ontologies. The 2 types of information (from events and ontologies) are complementary: event structures can determine what type of entities we have to look for in the ontology while the ontological information can disambiguate event types. The linguistic data they provide constitute a good basis for implementing strategies to improve video analysis.

2.5 Analysis of Complementary Textual Sources for adding Semantic Metadata to Multimedia Content

K-Space will reuse and implement mining methods and tools for the extraction of semantic features from complementary textual resources. Two different types of resources are considered:

- Mining and analysing primary resources: Analysis of the primary resources that are attached to the multimedia data, e.g. texts around pictures, subtitles of movies, etc. This can include textual information encoded in the textual annotation scheme of MPEG-7, and which is not generated from speech.
- Mining and analysis of secondary and tertiary resources: Analysis of data and text related to the multimedia data under consideration, e.g. a programme guide for a TV broadcaster or a web site displaying similar pictures.

We report here on a first experiment on the first kind of data, made within the past Esperonto project ([18]), and which gives a good basis for the work in K-Space. The topic of the small experiment described below was art, and the goal was to annotate semantically images of artworks in the web, on the base of surrounding texts and a small ontology on paintings. In this ontology, typical terms were associated to every class (so for example the terms "surrealism" and "cubism" are associated to the class "artistic-movement".

In the Esperonto scenario, we first defined the possibly relevant text regions for the semantic annotation of reproduction of paintings in web pages. Looking at an example showing a painting by Joan Miro, we identified following text regions (in both the text and in the html code) that are providing some descriptions of the painting:

1) Title of the document
2) Caption text: „Click on the image to enlarge" (a non relevant item, to be filtered by the tools, also on the base of lexical properties of the words).
3) Content of the HTML „Alt" tag: "'VEGETABLE GARDEN WITH DONKEY'"
4) Content of the HTML „Src" tag: http://www.spanisharts.com/reinasofia/miro/burrolt.jpg
5) Abstract text
6) Running text

On the base of this classification, we wrote a tool that supports the manual selection of such textual regions, and send those to a linguistic processing engine. The linguistic processing engine has been augmented with metadata specifying the type of text to be processed (we expect for example the Title and the "Alt" text to consist mostly of phrases.)

2.5.1 The Linguistic Annotation

In the following lines, we show some of the (partial) results of the linguistic analysis, as applied to the various text segments. Our tools are delivering a linguistic dependency annotation:

„Alt" text: 'VEGETABLE GARDEN WITH DONKEY';

```
<NP HEAD="garden" PRE_MOD="vegetable"
   <POST_MOD CAT= "PP" HEAD="with"
   NP_COMP_HEAD="donkey"</POST_MOD>
</NP>
```

Abstract/Running text: "…This picture depicts the rural landcape of Montroig …"

```
<SENT SUBJ="This picture" PRED="depicts
OBJ="the rural lansdscape of Montroig"</SENT>
```

Detailed annotation of the direct object "the rural landscape of Montroig":

```
<NP HEAD="landscape" PRE_MOD="rural"
<POST_MOD CAT="PP" HEAD="of"
NP_COMP_HEAD="Montroig"</POST_MOD>
</NP>
```

On the base of a mapping between the linguistic dependency and the terms associated to the classes of the (toy) ontology at hand, (whereas we augmented the classes of the ontology to be associated with patterns (for coping for example with date expressions), we could provide for a semantic annotation of the texts associated with the picture.

2.5.2 The (Toy) Art Ontology (schematized)

- Object > Artork > Painting [has_creator, has_name, has_subject, has_dimension, has_material, has_genre, has_date...]
- Person > Artist > Painter [has_name, has_birth_date, part_of_artistic_movement

2.5.3 The Instantiation of Classes

Mapping the output of the linguistic analysis onto the ontology allows to instantiate classes of the ontology with information extracted from the text:

1) Title: Vegetable garden with donkey
2) Creator: Miro
3) Date: 1918
4) Genre: naïve (if correctly extracted by some reasoning on the linguistically and semantically annotated text)
5) Subject: rural landscape of Montroig + garden and donkey (if the association between the title and the explanation given by the art expert can be grouped).
6) Dimension: 65x71
7) Material: Oil on canvas

In the experiment described above there was no principled relation between the terms in the ontology and the results of the image analysis (in term of low-level features). We think here that a domain ontology taking into account the specific features for the multi-modal analysis components could help in establishing this relationship. But clearly one has to think first of a principled integration of linguistic items in terms of possible indices of multimedia content, taking into consideration also the low-level features. This aspect builds one of the central research goals of K-Space.

3 Using Image Analysis for Improving Textual Information Extraction and Vice Versa

The relationship between the two types of information given in image and in text can also be exploited in an inverse mode: the properties of images have influence on the nature of text to be extracted. A typical example of scenario where the text is crucial while the images play an auxiliary role is the extraction of product information from web catalogues. In the Rainbow project [7,25], catalogues of bicycle products and of computer products were analysed by a tool based on statistical models (Hidden Markov Models) and simple 'wrapper' ontologies containing e.g. information about the data type and cardinality of various properties. Names, prices and other parameters of products were extracted together with

accompanying images. Images represent the context for text and vice versa.

In a baseline approach, all images were treated as equal for extraction purposes. In the next experiment, the images were pre-classified using three complementary methods: colour histogram analysis, image size analysis and latent semantic analysis (measuring the similarity of image bitmaps to pre-selected etalons). As expected, the combined use of both types of information improved the quality of extraction (significantly for images, less so for other types for information).

4 Conclusions

We have described some approaches that take advantages of the so-called complementary sources (text/transcripts) for automatically adding semantic metadata to image material. Till now we concentrated only on linguistic processing aspects, not taking into account low-level features extracted from image analysis. K-Space is offering a framework for integration of features extracted from all possible modalities at the knowledge level.

Acknowledgements

The research described in this paper is supported in part by the European Commission, contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content - K-Space, and in part by the SmartWeb project, which is funded by the German Ministry of Education and Research under grant 01 IMD01 A.

References

[1] T. Athanasiadis , V. Tzouvaras V., K. Petridis., F. Precioso, Y. Avrithis, Y. Kompatsiaris.. "Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content". In *proceedings of the ISWC 2005 Workshop "SemAnnot"* (2005)

[2] P. Buitelaar., M. Sintek., M. Kiesel. „Feature Representation for Cross-Lingual, Cross Media Semantic Web Applications". In: *Proceedings of the ISWC 2005 Workshop "SemAnnot"*, (2005)

[3] C. Clavel, T. Ehrette and G. Richard. "Events Detection for An Audio-Based Surveillance System". *International Conference on Multimedia and Expo* (IEEE-ICME'05), Amsterdam, The Netherlands. (2005)

[4] T. Declerck, J. Kuper, H. Saggion, A. Samiotou, P. Wittenburg, J. Contreras. "Contribution of NLP to the Content Indexing of Multimedia Documents". In *Lecture Notes in Computer Science* Volume 3115 / 2004 Pages 610-618, Springer-Verlag Heidelberg, 6 2004.

[5] J. Hunter: "Enhancing the semantic interoperability of multimedia through a core ontology." *IEEE Trans.Circuits Syst. Video Techn.* 13(1): 49-58 (2003)

[6] J. Hunter: "Adding Multimedia to the Semantic Web: Building an MPEG-7 ontology." *SWWS 2001*: 261-283 (2001)

[7] M. Labsky., P. Praks, V. Svatek V., O. Svab O.: "Multimedia information extraction from HTML product catalogues". In: *Workshop on Databases, Texts, Specifications and Objects* (DATESO'05), Ostrava 2005.

[8] A. Moreno , B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, J. Allen. "SPEECH DAT CAR. A Large Speech Database For Automotive Environments", *Proc. of LREC* 2000, Athens, (2000.)

[9] M. Alcántara: "Introduccion al analisis de estructuras linguisticas en corpus". In press, 2006.

[10] T. Brant: "TnT - A Statistical Part-of-Speech Tagger". In: Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000.

[11] T. Declerck: A set of tools for integrating linguistic and non-linguistic information. In: Proceedings of SAAKM (ECAI Workshop), 2002.

[12] W. Hsuy, Sh. Changy, Ch. Huangy, L. Kennedy, Ch. Linz, and G. Iyengarx: "Discovery and Fusion of Salient Multi-modal Features towards News Story Segmentation". In: SPIE/Electronic Imaging 2004, Jan. 18-22, San Jose.

[13] A. Merlino, D. Morey and M. Maybury: "Broadcast News Navigation using Story Segmentation". In: ACM Multimedia, Seattle, 1997, pp. 381-391.

[14] D. Oberle, A. Ankolekar, P. Hitzler, P. Cimiano, C. Schmidt, M. Weiten, B. Loos, R. Porzel, H.-P. Zorn, V. 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.

[15] Paul Buitelaar, Philipp Cimiano, Stefania Racioppa *Ontology-based Information Extraction with SOBA*. In: Proc. of LREC06, Genoa, Italy, May 2006

[16] Labský M., Svátek V., Šváb O., Praks P., Krátký M., Snášel V.: Information Extraction from HTML Product Catalogues: from Source Code and Images to RDF. In: 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), IEEE Computer Science, 2005.

Cited Projects:

[16] AceMedia project: http://www.acemedia.org/aceMedia
[17] BUSMAN project: http://busman.elec.qmul.ac.uk/
[18] Esperonto Project: http://www.esperonto.net
[19] K-Space Project: http://kspace.qmul.net
[20] SmartWeb Project: http://www.smartweb-projekt.de
[21] TRECVid: http://www-nlpir.nist.gov/projects/trecvid/
[23]MPEG-7: http://www.iso.org/iso/en/prods-services/popstds/mpeg.html
[24] MESH: http://www.mesh-ip.eu/
[25] Rainbow: http://rainbow.vse.cz