Enriching Videos with Light Semantics

Smitashree Choudhury
Digital Enterprise Research Institute
National University of Ireland, Galway
Galway, Ireland
smitashree.choudhury@deri.org

John G. Breslin
School of Engineering and Informatics
National University of Ireland, Galway
Galway, Ireland
john.breslin@nuigalway.ie

***Abstract*—This paper describes an ongoing prototypical framework to annotate and retrieve web videos with light semantics. The proposed framework reuses many existing vocabularies along with a video model. The knowledge is captured from three different information spaces (media content, context, document). We also describe ways to extract the semantic content descriptions from the existing user-generated content using multiple approaches of linguistic processing and Named Entity Recognition, which are later identified with DBpedia resources to establish meanings for the tags. Finally, the implemented prototype is described with multiple search interfaces and retrieval processes. Evaluation on semantic enrichment shows a considerable (50% of videos) improvement in content description.**

Keywords - social media; multimedia semantics; semantic web; linked open data; semantic search

I. INTRODUCTION

With the huge increase of user videos on the Web, the traditional search paradigm is proving to be ineffective in discovering and browsing interesting videos. Moreover, due to the complex nature of multimedia, reusability of video documents is very low, and as a result, almost every time a user has to create their video from scratch. We need better mechanisms to organise and represent the video data in order to address the above issues. Meaningful organisation and metadata representation is one of the concerns, but is as yet largely overlooked for multimedia. At present, user videos may come with certain embedded metadata, either created by users while publishing or during the production workflow, such as camera settings (though these are still not easily accessible in the case of web video). Some of the useful meta information is also created in the course of usage and sharing amongst users after publishing. Information such as free labels as tags, descriptions, user responses to the video, location information, membership in various groups, captions inside the video are immensely useful. The problem with the existing situation is that even if we collect and process this information, reusability (the data integration problem) remains elusive because of the lack of any formal semantics attached to the videos. Tags are freeform words with implicit meaning and relations known to the creator or publisher. The problems of user tagging have been explored well in many research studies. The major challenges are as follows. (1) Tag variation: different tags are used for the same kind of resources, e.g., "New York City", "NYC". There is no explicit way to express that these two tags are indeed meant to be the same. (2) Polysemy tags: a single tag used for different meanings. This problem occurs due to a difference in understanding of a user about the resource he or she is tagging, and may also depend on sociocultural differences among users. (3) Lack of formal structure among tags makes it difficult to understand, classify and recommend tags automatically. Besides these issues, we have problems with misspelling, compound tags such as "globalwarming", multiword tags expressed as multiple tags, and tags used out of a community consensus such as "SEMAPRO2010".

This plethora of information can be harnessed to add an extra layer of machine-readable metadata that will help to understand the opaque media data a little better. There are many well-defined and comprehensive formal ontologies available to describe media structures and content. The earliest such effort was made by the MPEG (Motion Picture Expert Group) community in developing MPEG-7 [7], a standard for describing media, but it failed to take hold significantly due to its lack of formal semantics and interoperability issues. The Semantic Web community made efforts [5] to convert MPEG-7 to RDFS (Resource Description Framework Schema) representations, in order to avail of the benefits offered by Semantic Web technologies such as RDF (the Resource Description Framework). However, the complexities of MPEG-7 prevented it from being fully converted and many data type issues remain unresolved. Media ontologies such as COMM [4] took a pure Semantic Web approach to describe and represent media with its different granularities. Many ontologies were developed to address domain-specific media such as museum collections, the football domain, etc.

Recently, the W3C Media Annotation Working Group has made an effort to devise a comprehensive media ontology to describe video on the Web, which may become a recommended standard in the near future. In spite of many concerted efforts, it is hard to see any widespread usage of these vocabularies. The reasons are not well studied, but on the other side we can see that there are some vocabularies such as FOAF (Friend of a Friend) [14], [20], SIOC (Semantically-Interlinked Online Communities) [13], which have been adopted quite well and quickly. We assume that the reasons for such adaptability may be due to their inherent simplicity and easy-to-understand characteristics. Keeping in mind the above challenges, we adopted the principle of keeping it short and simple (KISS), yet fulfilling the basic

 ISBN: 978-1-61208-104-5

requirements of ontology engineering, and proposed a lightweight framework to describe web videos. The approach makes use of many existing vocabularies such as Dublin Core, FOAF and SIOC wherever possible along with our own model. In spite of a very small and light framework, it covers almost every aspect of a media description. The description is broadly categorised under three sub modules: (a) document and media properties; (b) semantic content description; and (3) social context descriptions. Fig. 1 shows a subset of attributes from each of the three contributing information spaces. The details of the proposed model are in [12]. One of the focal points of the framework is its easy computability in the sense that most of the classes can be automatically populated with instances with little processing rules and heuristics. We have kept in mind the fact that in the future we may have to devise ways to map with the standard media ontology recommended by the W3C.

We also aim to link identified concepts to those of the Linked Open Data initiative (LOD), which was started in 2007 with the objective of creating a Web of Data connected to each other following four basic principles [11]. The hub of the Linked Open Data cloud is DBpedia, which is the RDF representation of Wikipedia [22] articles, categories and info boxes. Wikipedia is the largest user-generated multi-lingual encyclopedia in the world, maintained by tens of thousands of users since 2001. Other domain specific data sources such as book data, scientific publication data, life science data, geographical data are all connected to DBpedia [26] in the cloud. The present size of the LOD is more than 8 billion triples and is constantly increasing in size. More details of the LOD initiative can be found in [11].

The rest of this paper is structured as follows: in Section 2, we describe the related work. Section 3 describes the implementation flow including modeling, populating the model integration with linked data. Section 4 shows our semantic search prototype. Section 5 concludes this paper.

II. RELATED STUDIES

This section describes various studies related to semantic media modeling and semantic search of media data focusing on video search. It will also describe some efforts towards ontology learning from folksonomies. Ontology learning from folksonomies follows different approaches. Researchers in [6] suggested lightweight ontology learning from a folksonomy based on broader and narrower semantic relations. Passant [8] exploited folksonomies to populate a corporate ontology. Specia and Motta [10] used methods to cluster similar tags and find a match in an existing ontology.

Other studies proposed data mining technologies to mine the structural information from user tags. Schmitz et al. [9] used association rule mining techniques to recommend tags. Regarding semantic search, not much work has been carried out in the domain of multimedia data. A comprehensive study of semantic search is described in [1] while [2] describes an ontology-based search engine. A semantic video search system is described in [18]. Swoogle [17] and Sindice [3] are two major search engines focused on existing Semantic Web data.

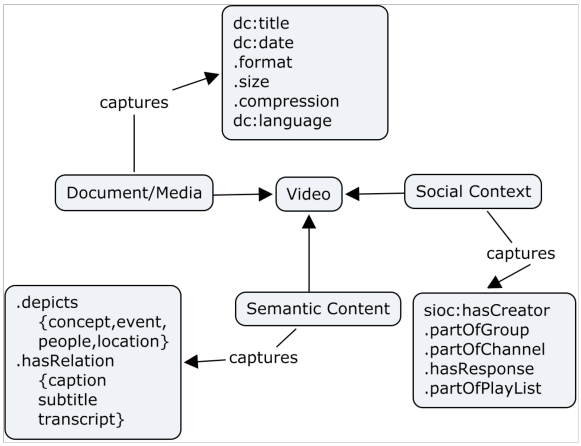

Figure 1. A subset of the video model.

III. IMPLEMENTATION ARCHITECTURE

This section describes various aspects of the prototypes including the instance creation, video annotation and retrieval modules. Fig. 2 shows the architectural flow of the prototype.

A. Data Collection

We used APIs and RSS feeds for different video sharing sites such as YouTube [23] and Vimeo [24] to collect the video metadata. Metadata includes title, description, tags, date, number of views, ratings, groups, duration, location data, etc. We have collected 10,000 video items for the prototype in the domain of science and technology.

B. Modelling Web Video

Our model for video description (Fig. 1) covers three major areas such as video document and media properties, social context attributes and depicted semantic content. The above proposed modeling approach not only satisfies the general ontological requirements such as modularity, interoperability and extensibility, but also separation of concern specifically aimed for media semantics. The uniqueness of the proposed approach for describing video is its simplicity and ease of use. Regarding the document level description, it is a widely-accepted practice to use Dublin Core terms such as title (*dc:title*) and creation date (*dc:date*), but media documents also carry some media-specific technical attributes such as format (*sva:format*), duration (*sva:duration*), etc., which are described using the video model described in [12]. Regarding the content description, video content can be described with different granularities starting with a global description (*dc:description*) to segments created by temporal and spatial decomposition. Segment content can be captured through the *sva:depicts* attribute whose range may be topic, event, geo-location, *foaf:Person* or *skos:Concept* as per requirements. The recent growth of social media interaction on the Web has made all objects on the Web somewhat social, thus we can embed some emerging properties such as comments, ratings, group membership, etc. For describing social contextual properties, the best-suited vocabulary is SIOC ontology. Its goal is to

 ISBN: 978-1-61208-104-5

describe objects and interactions in online communities. We consider the publisher of a video as an instance of *sioc:UserAccount* which belongs to a *foaf:Peson*. Video is an item in a *sva:Channel* which is a subclass of *sioc:Container*.

C. Content Processing for Concept Learning

Any ontology-based knowledgebase requires the instances to be populated manually, semi-automatically or by automatic means. Since manual annotation is not feasible and scalable, we tried to accomplish this semi-automatically by exploiting the existing information and getting user feedback in case of higher uncertainties such as the absence of any user data. APIs and RSS feeds offer an easy-to-go solution for many of the document level properties such as title, description, duration, categories, etc. which can be directly transformed to the Dublin Core properties or other global properties, but the real challenges come while creating the content description instances. The user-generated content is free text, devoid of any formal structure. In order to achieve the implicit formal structure, the content needed to be processed and normalised with various approaches before being mapped to any kind of ontological concepts.

Pre-processing of textual data involves:

- removing stop words
- removing tags with less than two characters
- removing username tags

After basic pre-processing we followed a few more intensive cleaning tasks in order to get some sensible tags from the data.

Multi-Term Tags: Tags with multiple words are one of the other major problems while identifying semantic entities. Mostly users enter multiple words as part of a single tag, and each of the tags are supposed to be separated by a comma delimiter, but the API gives a single word as a single tag. Taking the same example used previously, in many cases the YouTube API gives "global" and "warming" as two different tags while a single tag of "global warming" is more descriptive and accurate. In order to clean the tag space further and in the hope of getting some phrase tags, we followed a few simple syntactic rules (shown below) to parse the tag space. Examples of such rules are widely used in natural language processing research. After identifying the patterns, we check the resulting phrase with Wikipedia concepts, and if a match is found we keep the phrase as a possible candidate for a tag.

((Noun)+(Noun)*) or (Noun-Prep)?+(Adj|Noun)*

TABLE I. EXAMPLE OF MULTI TERM TAG IDENTIFICATION

Original tag space	Identified multi-term tags
sequencing, dna, rna, sanger, gilbert, big, dyes, terminators, molecular, biology, genomics, secuenciacin, adn, cidos, nucleicos	sequencing, dna, rna, sanger, gilbert, big, dyes, terminators, molecular biology, genomics

Entity Recognition (NER) with Open Calais: Open Calais [27] is a free non-commercial web service from Thomson Reuters for identifying various semantic entities such as person, event, location, company, dates, organisations, concepts, etc. Though its application is aimed at well-formed textual documents, we have tried it on tag spaces and description content as an experiment. The effectiveness of NER in tag spaces is expected to be lower because tags are independent words without any syntactic structure and grammar rules, but we assume that with careful cleaning and normalisation, we may be able to identify some entities. At present, entities identified from the tag space are only accepted if they are supported from other sources. When the video has more description content, use of Open Calais improves the result. Table II below shows five different identified entities from a video description.

TABLE II. EXAMPLE OF ENTITY IDENTIFICATION

Description content	Identified entities
Thus far, most DNA sequencing has been performed using the chain termination method developed by Frederick Sanger. This technique was also used to sequence the genome of James Watson recently. Pathogens may lead to treatments for contagious diseases. Biotechnology is a burgeoning discipline…	Contagious diseases Frederick Sanger (Person) James Watson (Person) Biotechnology (tech) DNA sequencing (tech)

Compound Tags: Users create tags with no white space, e.g., "globalwarming" which is a concatenation of two words "global" and warming". These tags are useful, but not in their original form, so we need to process them in order to separate the words with a whitespace and form a proper tag. We followed a few simple heuristics to identify meaningful words from a tag. The pseudo code is given below.

- *Divide the tag (T_i) into two sub tags (t_1, t_2) where length of t_1 is length((T_i)/2)+1 and t_2 is length((T_i)/2)-1*
- *Check if t_1 exists in the dictionary*
- *If(t_1 exists) = true*
  - *Check if t_2 exists*
  - *Form tag with t_1+WS+t_2 (equation 1)*
- *Else*
  - *Offset t_1 or t_2 with one character and check*
  - *If (one exists) then concatenate the offset and check if the other exists*
    - *Form the tag with t_1+WS+t_2 (equation 2)*
  - *Else (follow equation 3)*

Equation 3

If equation 2 fails, then we divide and create a third term t_3 with the offset characters and check iteratively. When two are found in the dictionary, we add the third by default and form the tag by adding a WS in between the terms. Though this is a brute force method it gives a satisfactory result for improving the tag quality

We restricted compound tags to a maximum of three terms. An example of the above algorithm is given below in Table III.

TABLE III. EXAMPLE OF COMPOUND TAG DECOMPOSITION

Original tag ("globalwarming")
Step 1. globalw (= t_1) and arming (= t_2)
Step 2. If (globalw is present in dictionary) = no
Step 3. Offset by 1 from t_1 (globalw-w = global) and add to t_2 (w+arming=warming)
Step 4. Check if t_1 and t_2 exists in dictionary = yes
Step 5. Form tag $T_i = t_1+WS+t_2$= **global warming**

D. *Integrating with the Linked Open Data (LOD) Cloud*

A video can be interlinked with multiple data sources such as geographical data, a *foaf:Person* or DBpedia concepts. Instances of concept, person, event, location are mapped with the property *owl:sameAs* or *rdfs:seeAlso*.

The focus here will be on content linking, from a tag to a Wikipedia concept to a DBpedia resource, e.g., the tag "E.coli" is mapped to a Wikipedia concept "Escherichia coli" and subsequently to the DBpedia resource "http://dbpedia.org/resource/Escherichia_coli". DBpedia is the hub of the LOD cloud, so any mapping to DBpedia will ultimately lead to other domain-specific data such as life science data or movie data.

Since there may not always be a one-to-one mapping between a user tag and an ontological concept, we need some kind of entity resolution mechanism. Here we computed a similarity between user tags and wiki concepts (from wiki articles) and redirect concepts, and derive the top match as the identified concept. This particular similarity is computed with a Lucene index of Wikipedia articles, redirects and categories.

E. *Semantic Relation Extraction*

Once we get a list of probable tags from all of the above steps, we need to formally ground them with some ontological concepts with relations between them. Since at all stages in the above processing we verified the possible tag against an index of Wikipedia articles, categories and redirect concepts, they are more or less considered ontological concepts though the relationship between them is still unclear and vague.

To extract the relationship between tags we need to compute the similarity between tags. Many studies explored tag similarity using various approaches and distributional measures such as co-occurrence similarity [16], Folkrank [15], etc. At the time of writing, this similarity module has not been implemented, but we plan to exploit the link structure of Wikipedia articles to estimate the semantic distance between tags.

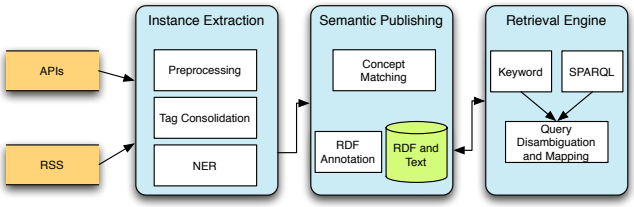

Figure 2. System architecture modules.

IV. SEARCH MODULE

Machine-readable data will facilitate complex query answering which was not possible before. It will also help to infer some unseen relations existing between various data pieces within the knowledgebase (KB) itself, but it still remains insulated from the huge amount of data lying outside the KB which may hold much more relevant and useful information both known and unknown.

Here come the benefits of linked data: by following some simple principles we can make our data accessible to other datasets and vice versa. The benefits of linked data can only be realised with practical applications, so we have decided to enable our semantic search module to explore the linked data to facilitate navigational search, where the user can explore and discover much related information and therefore reformulate their queries. Fig. 3 shows an interface for the query "Albert Einstein", and its related information as aggregated from the DBpedia source.

A. *User Interface*

The role of good user interfaces for Semantic Web data has largely been overlooked. To our understanding, it is one of the major contributing factors to the slow adoption of Semantic Web technologies. Although recently some efforts have been made to address the issue, such as faceted browsers like mSpace [19] and Sigma [21], the problem is far from over.

The ideal solution should not reflect underlying data complexities but still give the benefits of semantic search. [22] is a standard recommendation for querying Semantic Web data, but exposing a SPARQL interface as the primary query interface will be riskier as learning a complex query language will hardly be welcomed by users other than concerned geeks. A simple keyword-based interface may suffice for most users, but will lose the complex query answering mechanisms possible with semantic data.

Therefore we have planned to expose different levels for a query interface in order to facilitate complex queries by exposing underlying data properties with each querying stage. We move from keyword search to faceted search, where the major facets are dynamically constrained for each iteration, and finally to navigational search. Navigational search enables the user to access an integrated view of the query term. Fig. 3 shows the incremental query interface of the system. The first point of entry is a dual interface of keyword search and SPARQL end point. The result of the first query is deployed in a faceted interface. Details of the video are exposed in a navigational space where related facts are connected DBpedia resources.

 ISBN: 978-1-61208-104-5

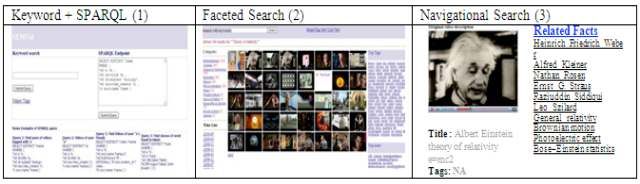

Figure 3. Three different interfaces.

B. *Retrieval Engine Architecture*

Since semantic search is defined as the matching between query semantics and content semantics, we need to capture the query semantics before the actual search process. User query intention can be captured in different ways starting from interacting with query reformulation to automatic disambiguation of a query.

- For the keyword search interface, we have adopted a simple approach to disambiguate the user queries by mapping the query term(s) to the best possible semantic entity that exists in the knowledgebase. In the case of more than one semantic entity, entity resolution is performed in favour of the most popular one, followed by the rest. However, in such cases, precision goes down. We need to adopt a more robust entity resolution mechanism in order to improve the search quality.
- At the second stage, the query is sent to the Lucene index for retrieval. The results are clustered with various facets such as top-related tags in the result set, top categories, top users for the query, dominant timeline, etc.
- On the faceted interface, the user can get a glimpse of the underlying data attributes and can filter the result with each iteration.
- Clicking on a single thumbnail will lead to a video detail page (navigational search) where the video is displayed not only with the original descriptions, but also with some extra resources related to the user query concept.
- These resources are connected to the user query concept. There may be too many resources in one DBpedia page and all are not of equal relevance. In future, we need to figure out how to rank the connected resources in relation to the query concept. One heuristic may be to rank the resources of a similar type higher compared to the others, or we can compute a resource distance based on mutual information sharing such as categories, property values, etc. This part of the work is still ongoing.

V. EVALUATION

Since the evaluation is still ongoing at the time of writing, we report a part of the evaluation. The objective here is to evaluate the effectiveness of the automatic augmentation of light semantics from various sources and its impact on retrieval in terms of user satisfaction.

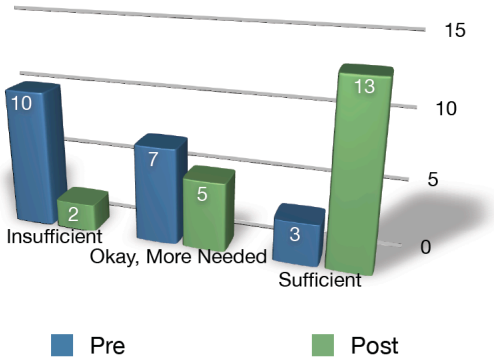

Figure 4. Evaluation of task 1 (content enrichment).

Effectiveness and user satisfaction are both measured qualitatively based on user ratings. Five users evaluated 20 random videos for their content description sufficiency. Each user was presented with a list of inferred keywords for describing the video content and were asked to rate the list for degree of sufficiency on a three-point scale of 1 to 3, after watching the video. The average video duration in the evaluation was 3.25 minutes.

A rating of 1 is the least descriptive (insufficient or irrelevant), while 3 is rated as a sufficient description of the depicted content, and a rating of 2 is considered as representing that there are some descriptions but more are needed. The result is based on inter-user agreement of ratings (a minimum of 3 out of 5 users agreed for a score).

Figure 4 shows the results of the evaluation of task 1, where the number of sufficient content descriptions increases to 13 videos from only 3 videos, whereas 5 videos are still considered to be in need of more descriptive keywords. The average rating per video increased from 1.65 to 2.5. In the evaluation of task 2, we have started to measure the level of user satisfaction for search results after enrichment.

VI. CONCLUSIONS AND FUTURE WORK

We have discussed a lightweight framework to provide metadata for user videos on the Web using several existing ontologies. We discussed an approach to create instance data based on our models from user-generated content using both statistical and linguistic approaches.

We also described our approach to integrate the structured video data into the Linked Open Data cloud for greater integration and interoperability. Finally, the paper details an implemented prototype for the semantic search of web videos with three different modes of user interface.

Our future work involves robust evaluation of the instance-learning module and the creation of a fully-fledged integrated semantic annotation and search system.

ACKNOWLEDGMENT

This work was supported by Science Foundation Ireland under grant number SFI/08/CE/I1380 (Líon 2).

 ISBN: 978-1-61208-104-5

REFERENCES

[1] C. Mangold, "A survey and classification of semantic search approaches", Int. J. Metadata, Semantics and Ontology, vol. 2, pp. 23-34, 2007.

[2] D. T. Tran, S. Bloehdorn, P. Cimiano, and P. Haase. "Expressive resource descriptions for ontology-based information retrieval", Proc. of the 1st Int. Conf. on the Theory of Information Retrieval (ICTIR '07), October 2007.

[3] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello, "Sindice.com: a document-oriented lookup index for open linked data", IJMSO, vol. 3, no. 1, pp. 37-52, 2008.

[4] R. Arndt, R. Troncy, S. Staab, L. Hardman, and M. Vacura, "COMM: Designing a Well-Founded Multimedia Ontology for the Web", Proc. of the 6th International Semantic Web Conference (ISWC 2007), Busan, Korea, November 11-15, 2007.

[5] J. Hunter, "Adding multimedia to the Semantic Web – Building an MPEG-7 ontology", 1st International Semantic Web Working Symposium (SWWS '01). California, USA, pp. 261–281, 2001.

[6] P. Mika, "Ontologies are us: a unified model of social networks and semantics", ISWC 2005, LNCS vol. 3729, pp. 522–536, Springer, 2005.

[7] MPEG-7: Multimedia Content Description Interface, ISO/IEC 15938, 2001.

[8] A. Passant, "Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs", International Conference on Weblogs and Social Media, 2007.

[9] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme, "Mining association rules in folksonomies", Data Science and Classification, pp. 261–270, 2006.

[10] L. Specia, E. Motta, "Integrating folksonomies with the semantic web", Proc. of the 4th European Conference on the Semantic Web: Research and Applications, Innsbruck, Austria, 2007.

[11] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, "Linked data on the Web," Proc. of the 17th Int. Conf. on World Wide Web, ACM, 2008, pp. 1265–1266.

[12] S. Choudhury, J. Breslin, and S. Decker, "A lightweight web video model with content and context descriptions for integration with linked data", Proc. of Semantic Authoring, Annotation and Knowledge Markup Workshop, 2009.

[13] J.G. Breslin, A. Harth, U. Bojars, and S. Decker, "Towards semantically-interlinked online communities", Proc. of the 2nd European Semantic Web Conference (ESWC '05), LNCS vol. 3532, pp. 500-514, Heraklion, Greece, 2005.

[14] D. Brickley, L. Miller, "The Friend Of A Friend (FOAF) vocabulary specification", http://xmlns.com/foaf/0.1/, 2005.

[15] C. Cattuto, D. Benz, A. Hotho, and G. Stumme, "Semantic grounding of tag relatedness in social bookmarking systems", Proc. of the 7th International Semantic Web Conference, 2008.

[16] H. Halpin, V. Robu, and H. Shepherd, "The complex dynamics of collaborative tagging", WWW 2007, ACM Press, 2007.

[17] L. Ding, T. Finin, A. Joshi, R. Pan, R.S. Cost, Y. Peng, P. Reddivari, V.C. Doshi, and J. Sachs, "Swoogle: A search and metadata engine for the semantic web", Proc. of the 13th ACM Conference on Information and Knowledge Management, 2004.

[18] J. Waitelonis, H. Sack, "Augmenting video search with linked open data", Proc. of Int. Conf. on Semantic Systems, 2009.

[19] M. Schraefel, A. Wilson, A. Russell, and D. A. Smith, "mSpace: improving information access to multimedia domains with multimodal exploratory search", Commun. ACM, vol. 49, no. 4, pp. 47–49, 2006.

[20] E. Prud'hommeaux, A. Seaborne, "SPARQL query language for RDF", W3C, 2008.

[21] Sigma: Available online at http://sig.ma

[22] Wikipedia: Available online at http://www.wikipedia.org

[23] YouTube: Available online at http://www.youtube.com

[24] Vimeo: Available online at http://www.vimeo.com

[25] W3C Media Annotation Group: Available online at http://www.w3.org/2008/WebVideo/Annotations/

[26] DBpedia: Available online at http://dbpedia.org

[27] OpenCalais: Available online at http://www.opencalais.com

 ISBN: 978-1-61208-104-5