published as: Waitelonis, J. & Sack, H.: Augmenting Video Search with Linked Open Data. In Proceedings of International Conference on Semantic Systems (i-semantics), September 2-4, 2009, Graz, Austria, Verlag der TU Graz, Austria, 2009. (ISBN 978-3-85125-060-2)

# Augmenting Video Search with Linked Open Data

**Jörg Waitelonis, Harald Sack**
(Hasso-Plattner-Institute Potsdam, Germany
{joerg.waitelonis|harald.sack}@hpi.uni-potsdam.de)

**Abstract:** Linked Open Data has become one of the driving forces for the emerging Semantic Web, which enables interlinking and integrating former proprietary data to the global linked data by using RDF as a standard data format. In this paper we show how to integrate the database of the open academic video search plattform yovisto.com with the linked open data cloud and how to augment yovisto video search by including semantically interrelated linked open data.

**Key Words:** semantic web, linked data, multimedia retrieval, video retrieval
**Category:** H.3.3, H.5.1, H.5.4

## 1 Introduction

Publication of audiovisual content in the World Wide Web (WWW) has never been so popular as the tremendous success of video sharing platforms such as YouTube or Google-Video is proving. ComScore announced that more than 147 million U.S. Internet users have watched an average of 101 videos per viewer in January 2009 [comScore, Inc., 2009]. Therefore searching, serializing, categorization, and filtering of web video are essential to assist the user to localize videos of their interests.

Yovisto.com[1] is a video search engine specialized on video recordings of academic lectures and conference recordings. Its search index provides efficient access to more than 6 200 content-based searchable lecture recordings from universities and scientific institutions all around the world. To enable pinpoint search access within the video recordings a combination of automated video analysis and user generated collaborative annotation is deployed in connection with fine-granular, time-dependent metadata [Sack and Waitelonis, 2006].

Videos available at yovisto are annotated with time-based MPEG-7 encoded metadata [Day and Martínez, 2000], which demands sophisticated search methods, to provide the user with a powerful tool to investigate the data. State-of-the-art keyword based search technologies are used to provide a fast and efficient access, but nevertheless the keyword based search experience is rather unidimensional. Although users might filter search results by categories, tags, and other facets, the scope of the achieved results is only narrowed with these techniques. If users are not able to specify what exactly they are looking for, they easily got

[1] http://www.yovisto.com/

lost in a funnel trying to focus the achieved search results. E. g., "more-like-this" queries are suggesting similar results giving a clue about other resources, which could be in their scope of interest. Determined by statistical clustering, these results usually are very similar to the primary result set, i. e. users are only moving within the same cluster or domain without the chance to look what is waiting outside. Social network based methods such as, e. g., collaborative tagging can help to broaden the search scope beyond similarities, e. g. “Users who bought X, also bought Y.”.

In this paper we will show how to deploy Semantic Web technologies to enrich search results of the yovisto video search engine and how to improve user experience by enabling a semantically supported explorative search. To achieve this, we have connected yovisto's content with the Linked Data web [Bizer et al., 2008] to incorporate external additional information (link-up) as well as to use this external information to uncover cross-connections back to yovisto's own content (link-back).

The rest of the paper is structured as follows. Section 2 refers to related work, while Section 3 introduces use cases that show how the video search data can be complemented with Linked Open Data. Section 4 details the realization of all steps to proceed the linkage, Section 5 points out briefly first experiences, results and first bits of evaluation. Section 6 concludes the paper.

## 2 Related Work

Linked Open Data (LOD) has become one of the most popular topics among the emerging Semantic Web [Berners-Lee, 2006]. By Linked Data we refer to a method of exposing, sharing, and connecting data via dereferenceable URI on the WWW. The Linking Open Data community project has picked up this approach to extend the Web of data by publishing various open data sets being represented as RDF and by defining links and mappings between vocabularies and data items from different data sources [Bizer et al., 2008]. Thus, the data available within the LOD has grown to more than 4.5 billion RDF triples and about 180 million RDF links[2]. This publicly available interconnected data enables the development of numerous data mash-up applications.

One of the most prominent datasets among LOD is the DBpedia that provides all structured data of the well known online encyclopedia Wikipedia[3] being represented as RDF triples [Auer et al., 2008]. DBpedia contains about 2.18 million concepts described by 218 million triples, including abstracts in 11 different languages.

[2] `http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics`, (March 2009)

[3] `http://www.wikipedia.org/`

Another dataset which is important for our work is the Digital Bibliography & Library Project (DBLP)[4]. DBLP is hosted at the University of Trier (Germany) and originally it was a database and logic programming bibliography site, which has broadened its scope to become one of the largest scientific bibliographical databases on the Web. The D2R Server at the L3S Research Center of the Leibniz University in Hannover (Germany) provides a weekly updated RDF-extract of DBLP's 950 000 scientific articles and 570 000 being encoded in about 28 million RDF triples[5].

GeoNames[6] is a geographical data base available and accessible through various Web services, under a Creative Commons attribution license. It contains more than 8 million geographical names corresponding to over 6.5 million unique features and additional data such as, e. g., latitude, longitude, elevation, or population. GeoNames data are linked to DBpedia data and other RDF Linked Data.

The Friend of a Friend (FOAF) ontology is describing persons, their activities and their relationships to other people and objects [Brickley and Miller, 2007]. With FOAF people are able to describe social networks without the need for a centralised database. People are publishing FOAF profiles as RDF encoded files on their homepages, where they can easily be found with the help of specialized Semantic Web search engines such as, e. g., Sindice [Tummarello et al., 2008].

## 3 Use Cases for Linked Data in Video Search

### 3.1 Complementing Video Data with Linked Open Data

For the first use case, we use Linked Open Data to complement already existing information about entities within the scope of the domain of our academic video search engine. Each of yovisto's video recordings features one or more speakers, who give a presentation about one or more specific scientific subjects. E. g. the resource "organisation" denotes the university or company to which a speaker belongs. By default, yovisto's database only provides the following properties for the organization: name, country, city, type (university or other), and website.

For universities, DBpedia provides numerous complementary information such as, e. g., an abstract summary, the motto of the university, the president, celebrities who studied or worked at the university, geographical information, and so forth. By identifying a university by its name and location, one can access the according DBpedia dataset and integrate all interesting information into yovisto automatically.

[4] `http://www.informatik.uni-trier.de/∼ley/db/`

[5] `http://dblp.l3s.de/d2r/`

[6] `http://www.geonames.org/`

To obtain scientific (and other) publications of a speaker, one can also use DBpedia data about the speaker, if unique identification is possible and if the speaker actually has a DBpedia entry. Otherwise, i. e. if the speaker is not prominent enough to have an DBpedia article (and Wikipedia article respectively), we might query the DBLP bibliography about the speakers publications. By using DBLP data, yovisto speaker data can be endorsed with bibliographical data about the speaker's publications automatically.

By including geographical information via DBpedia and GeoNames, locations of search results such as, e. g., the place where the lecture was recorded or the organization (university) where the speaker is employed can be arranged in a geographical map mash-up (cf. Fig. 1).

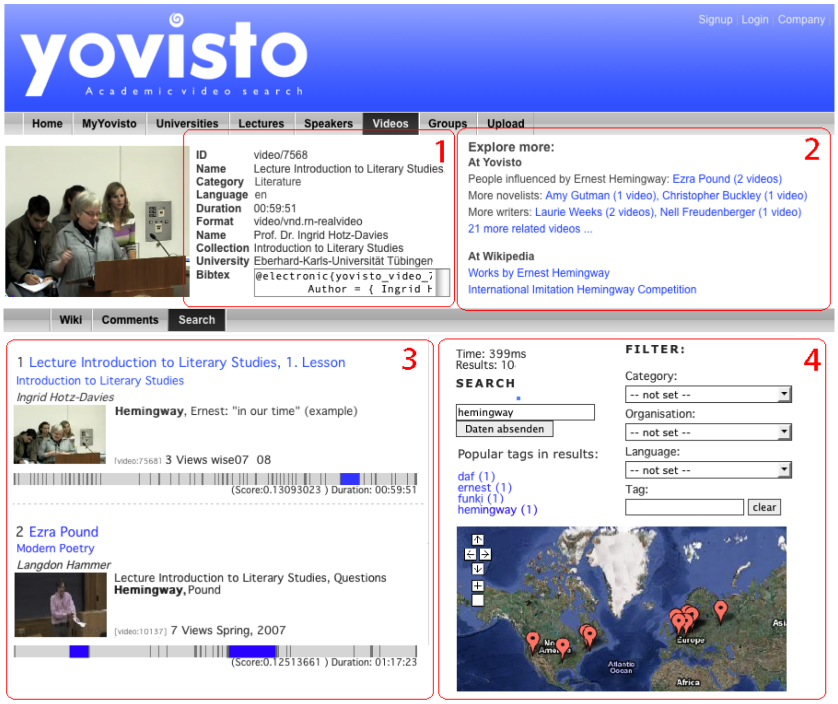
### 3.2 Enable Explorative Search with Linked Open Data

In the second use case we utilize Linked Open Data information to enable an explorative video search, i. e. related search results will be arranged in a way to guide the user on his expedition through yovisto's entire search space. Traditional search engines work on a strict input/output basis, i. e. the user inputs a search phrase and the search engine delivers a list of results ordered by their relevance. If the user changes the search phrase, the result list is also altered. By including or excluding new search terms and connecting them with Boolean connectors, the search scope can be narrowed or broadened. But, this way to explore the entire offering of a search engine is rather inefficient.

By connecting yovisto resources with Linked Open Data resources, previously implicit cross connections can be made explicit. This means to connect video resources amongst each other, which are not related at a first glance. E. g., suppose a search query for the american novelist "Ernest Hemingway". Via DBpedia we can find out that Hemingway refers to a person and that many other persons are related with Hemingway via the properties `dbpedia:influenced` and `dbpedia:influences`. By evaluating the property labels we can enrich the displayed result with links to other people's Wikipedia information pages labeled with "influenced (by)" and "influences", accordingly. We are not able to locate these interrelationships without the help of Linked Data. The following simple SPARQL query on DBpedia delivers a list of people influencing or being influenced by Ernest Hemingway:

```
SELECT ?label ?person WHERE {
     ?ernest foaf:name "Ernest Hemingway" .
     {?person dbpedia2:influenced ?ernest .
      dbpedia2:influenced rdfs:label ?label . }
    UNION
     {?ernest dbpedia2:influences ?person .
      dbpedia2:influences rdfs:label ?label . }
}
```

By passing the result as a list of speakers to yovisto, new video suggestions can be made that would not be possible by using yovisto data only. These newly created exploration paths are displayed within an additional widget as navigation element besides the regular video metadata on the yovisto website (cf. Fig. 1). In general, all places, person names, proper names of entities, or dates



| | |
|---|---|
| ID | video/7568 |
| Name | Lecture Introduction to Literary Studies |
| Category | Literature |
| Language | en |
| Duration | 00:59:51 |
| Format | video/vnd.rn-realvideo |
| Name | Prof. Dr. Ingrid Hotz-Davies |
| Collection | Introduction to Literary Studies |
| University | Eberhard-Karls-Universität Tübingen |
| Bibtex | @electronic{yovisto_video_7<br>Author = { Ingrid H |

**Explore more:**
**At Yovisto**
People influenced by Ernest Hemingway: Ezra Pound (2 videos)
More novelists: Amy Gutman (1 video), Christopher Buckley (1 video)
More writers: Laurie Weeks (2 videos), Nell Freudenberger (1 video)
21 more related videos ...

**At Wikipedia**
Works by Ernest Hemingway
International Imitation Hemingway Competition

1 Lecture Introduction to Literary Studies, 1. Lesson
Introduction to Literary Studies
*Ingrid Hotz-Davies*
**Hemingway**, Ernest: "in our time" (example)
[video:7568] 3 Views wise07 08
(Score:0.13093023 ) Duration: 00:59:51

2 Ezra Pound
Modern Poetry
*Langdon Hammer*
Lecture Introduction to Literary Studies, Questions
**Hemingway,** Pound
[video:10137] 7 Views Spring, 2007
(Score:0.12513661 ) Duration: 01:17:23

Time: 399ms
Results: 10
**SEARCH**
hemingway
Daten absenden
Popular tags in results:
daf (1)
ernest (1)
funki (1)
hemingway (1)

**FILTER:**
Category: -- not set --
Organisation: -- not set --
Language: -- not set --
Tag: clear

Figure 1: Search result for “hemingway”. Next to the standard display of video metadata (1) the exploration navigation (2) as extension to regular keyword based search (3) and search filters (4) is displayed. In addition, geographical data referring to organizations of the results set are visualized.

stemming from video segment keywords can be interlinked with DBpedia concepts. In fact, almost all entities are subsumed within a system of categories, which can be determined via the property `rdf:type`. Now, by retrieving other entities of the same category (type) gives way to an explorative search within the yovisto database by retrieving video segments being indexed with the name of these entities. Thus, video results related to the original search with regard to categories (super-classes) enable a new way to explore the yovisto search space. E. g., the following SPARQL query results in video segment identifiers that are related to entities of the same categories as dbpedia:Ernest_Hemingway, i. e. other writers or novelists:

```
SELECT DISTINCT ?item ?type ?segment WHERE {
    dbpedia:Ernest_Hemingway rdf:type ?type.
    ?item rdf:type ?type.
    ?segment rdf:type yovisto:videosegment.
    ?segment foaf:depicts ?item.
}
```

By connecting all occuring concepts within the yovisto search engine index data to the Linked Data space, we are contributing audiovisual data (i. e. video segments containing these concepts or being related with these concepts) to the Linked Open Data initiative.

## 4 Linked Data and Interlinking Concepts

This section briefly describes how the yovisto database has been interconnected to Linked Open Data. To establish yovisto as a linked data application we follow the principles and guidelines decribed in [Hausenblas, 2009]. Accordingly to the Linked Data concept, we gather external data first, weave it together with our own data, and publish the resulting linked data finally.

There are various points of contact to tie yovisto's data to other resources such as, e. g., organizations (DBpedia, GeoNames), persons (DBpedia, FOAF, Sindice), categories and scientific fields (DBpedia, OpenCyc) [OpenCyc, 2009], or publications (DBpedia, DBLP).

Interlinking with yovisto should be automated as much as possible. For that reason, we matching algorithms had to be implemented to map yovisto resource labels to the labels of concepts of the other linked data provider. Matching the yovisto top level categories was performed manually, because of the small number of about 30 items only. Matching organization and university labels was more difficult. Comparing names directly with all rdf:label subjects of DBpedia was not possible, because of performance issues. Therefore, we reduced the search space by restricting the dbpedia-owl:country to the country set in our database and rdf:type to dbpedia-owl:organisation.

Additionally, by removing stopwords and tokenizing the labels, regular expressions for every token have been constructed using the boolean AND operator to be invariant from the serialization of words (e. g., by querying for "Univ AND Yale" we can match "Univ. of Yale" well as "Yale University"). To match persons we utilized the sindice API to retrieve FOAF profiles and links to DBpedia resources. We used the combined search method of sindice API to query for given name, family name, and both with restriction to class:Person and foaf:name, foaf:givenname.

Besides the automated mapping to (already existing) resources from e. g. DBpedia an indirect manual mapping can be applied to reveal proper matchings. E.g., if a user is creating a new organization, which is not in the yovisto database, she will be assisted by automated suggestions in the course of a step-by-step navigation. When a user is creating a university that is not in the yovisto database, she has to choose the country first, then the city and finally, a list of universities for that particular city are displayed to choose from. For every step a suggestion list is generated from DBpedia (and/or Geonames).

Publishing the data means first to have a look on the non-RDF data structure and express this structure as taxonomy, create URIs for the resources (minting), choose existing description vocabularies, and publish the data via RDFa or a SPARQL endpoint.

Yovistos data consist of certain cascaded entities. A video (e. g. a recorded conference talk) is part of a collection (e. g. conference session, lecture, podcast, etc.). A collection belongs to an organization (e.g. university or conference). A video is related to one or more agents (e. g. speaker, presenter) and is subdivided into coherent video segments as regards content. Hence, a taxonomy can be defined almost straightforward and described with SKOS [Miles and Brickley, 2005, Summers et al., 2008].

Yovisto data being structured in a relational database and the Virtuoso [OpenLink, 2009] triple-store. Resources are identified via URIs such as `http://www.yovisto.com/resource/video/2359`. Depending on the client's content type acceptance, content negotiation redirects to either RDF-documents (if set to `application/rdf+xml`), or the HTML documents. Furthermore, `http://www.yovisto.com/resource/lecture/2142` dereferences a particular lecture instance with identifier 2142, and `http://www.yovisto.com/resource/lecture` dereferences the class lecture (in sense of yovisto), which has a link `skos:broader` to `http://www.yovisto.com/resource/collection` as a super-class and a `owl:sameAs` connection to `http://dbpedia.org/resource/Lecture` on DBpedia.

We use the following vocabularies to describe the yovisto resources: yovisto ontology[7], FOAF and VCARD [Iannella, 2001] for persons, SKOS to define taxonomies, DublinCore [International Organization for Standardization, 2003] for bibliographic metadata, LOM [Hodgins and Duval, 2002] for educational metadata. For the description of technical characteristics as well as the temporal segmentation of videos, yovisto deploys MPEG-7 metadata, which can be mapped to the Core Ontology for Multimedia (COMM) [Arndt et al., 2007].

Finally, the data get exposed via RDFa annotation being embedded in the human-digestable XHTML[Pemberton, 2002] pages as well as via a SPARQL endpoint[8] for public access implemented by Virtuoso.

[7] http://www.yovisto.com/ontology

[8] http://sparql.yovisto.com/sparql

## 5 Results and Brief Evaluation

This section gives a brief overview of a few first results and evaluations of our very first approach to connecting yovisto to Linked Data. There is no full confidence achievable when working with Linked Data. Therefore, automated solutions for mapping entities to Linked Data such as, e. g. DBpedia cannot guarantee correct and exhaustive results, because of the manifold and various application of categories and properties as regards content. For every usecase described in section 3 a set of queries had to be created to obtain valid results. Different queries have to be created for obtaining similar results regarding the content.

A secondary outcome of the work was to find out how many items of yovisto can be interlinked at all with this first (semi-automated) approach. From 1100 speakers we were able to interlink automatically 130 (12 %), with 5 (3.9 %) of them incorrect. 88 persons were linked to DBpedia, for the other 44 we found FOAF profiles via sindice to gather additional information.

From 440 organizations we were able to interlink 80 (18.2 %), with 5 (6.25 %) of them incorrect. The incorrect matchings occured most frequently on universities being represented in different DBpedia articles such as, e. g., "Columbia University" was likewise matched to "Columbia International University", "The School at Columbia University" , "University of the District of Columbia".

Smilar to Kobilarov et. al. who succeeded in linking 20 % to 30 % of BBC media brands, locations, names, and subject to DBpedia [Kobilarov et al., 2009], we observed too, that the most items don't have an article in Wikipedia, where to link to or an existing FOAF profile in case of persons.

## 6 Conclusion and future work

In this paper we showed how the video search engine yovisto has been augmented with Linked Data to improve user experience by enabling a semantically supported explorative search. We used Linked Data to complement yovisto video data and to uncover implicit cross connections within our data. We discussed, how we have realized the interlinking through gathering of Linked Data and mapping our concepts to Linked Data, to finally publish the results embedded in RDFa and a SPARQL endpoint.

Not all problems could be discussed in detail and there are open issues to solve in future work. Matching single keywords to concepts in DBpedia needs disambiguation methods, which have not been discussed. Preparation of high frequent queries beforehand, caching and pre-indexing workflows have to be discussed in future work. Furthermore, the concept of explorative search being introduced in this paper can be generalized to traditional WWW search engines.

## References

[Arndt et al., 2007] Arndt, R., Troncy, R., Staab, S., Hardman, L., and Vacura, M. (2007). COMM: Designing a Well-Founded Multimedia Ontology for the Web. In *ISWC 2007 + ASWC 2007*, pages 30–43.

[Auer et al., 2008] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2008). DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007)*, pages 722–735.

[Berners-Lee, 2006] Berners-Lee, T. (2006). Linked data. World wide web design issues.

[Bizer et al., 2008] Bizer, C., Heath, T., Idehen, K., and Berners-Lee, T. (2008). Linked data on the web. In Huai, J., Chen, R., Hon, H.-W., Liu, Y., Ma, W.-Y., Tomkins, A., and Zhang, X., editors, *Proceedings of the 17th International Conference on World Wide Web (WWW)*, pages 1265–1266. ACM.

[Brickley and Miller, 2007] Brickley, D. and Miller, L. (2007). The Friend Of A Friend (FOAF) vocabulary specification.

[comScore, Inc., 2009] comScore, Inc. (2009). Press release: Youtube surpasses 100 million u.s. viewers for the first time.

[Day and Martínez, 2000] Day, N. and Martínez, J. M. (2000). Introduction to MPEG-7. Technical Report ISO/IECT JTC1/SC29/WG11 N3751, International Organisation for Standardisation.

[Hausenblas, 2009] Hausenblas, M. (2009). Exploiting Linked Data For Building Web Applications. *IEEE Internet Computing*, N(N):N–N.

[Hodgins and Duval, 2002] Hodgins, W. and Duval, E. (2002). Draft standard for learning technology - Learning Object Metadata - ISO/IEC 11404. Technical report, ISO/IEC.

[Iannella, 2001] Iannella, R. (2001). Representing vCard Objects in RDF/XML. World Wide Web Consortium, Note NOTE-vcard-rdf-20010222.

[International Organization for Standardization, 2003] International Organization for Standardization (2003). Information and Documentation – The Dublin Core Metadata Element Set. ISO 15836.

[Kobilarov et al., 2009] Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., and Lee, R. (2009). Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Conections. In *European Semantic Web Conference, Semantic Web in Use Track, Crete.*

[Miles and Brickley, 2005] Miles, A. and Brickley, D. (2005). Skos core vocabulary specification. World Wide Web Consortium, Working Draft.

[OpenCyc, 2009] OpenCyc (2009). http://opencyc.org/.

[OpenLink, 2009] OpenLink (2009). Virtuoso http://virtuoso.openlinksw.com/.

[Pemberton, 2002] Pemberton, S. (2002). XHTML 1.0: The Extensible HyperText Markup Language (Second Edition). World Wide Web Consortium, Recommendation REC-xhtml1-20020801.

[Sack and Waitelonis, 2006] Sack, H. and Waitelonis, J. (2006). Integrating Social Tagging and Document Annotation for Content-Based Search in Multimedia Data. In *n Proc. of the 1st Semantic Authoring and Annotation Workshop (SAAW2006)*, Athens (GA), USA.

[Summers et al., 2008] Summers, E., Isaac, A., Redding, C., and Krech, D. (2008). LCSH, SKOS and Linked Data. *CoRR.*

[Tummarello et al., 2008] Tummarello, G., Delbru, R., and Oren, E. (2008). Sindice.com: Weaving the Open Linked Data. *The Semantic Web*, pages 552–565.