SemWebVid - Making Video a First Class Semantic Web Citizen and a First Class Web Bourgeois*

(Submission to the Open Track of the Semantic Web Challenge 2010)

Thomas Steiner#, Michael Hausenblas§,

#Google Germany GmbH, ABC-Straße 19, 20354 Hamburg, Germany
`tsteiner@{google.com, lsi.upc.edu}`¶
§National University of Ireland, DERI, IDA Business Park, Lower Dangan, Galway, Ireland
`michael.hausenblas@deri.org`

Abstract. SemWebVid[1] is an online Ajax application that allows for the automatic generation of Resource Description Framework (RDF) video descriptions. These descriptions are based on two pillars: first, on a combination of user-generated metadata such as title, summary, and tags; and second, on closed captions which can be user-generated, or be auto-generated via speech recognition. The plaintext contents of both pillars are being analyzed using multiple Natural Language Processing (NLP) Web services in parallel whose results are then merged and where possible matched back to concepts in the sense of Linking Open Data (LOD). The final result is a deep-linkable RDF description of the video, and a "scroll-along" view of the video as an example of video visualization formats.

Keywords: RDF, LOD, Linked Data, Semantic Web, NLP, Video

1 Introduction

Over recent years the use of Resource Description Framework (RDF) in documents has gained massive popularity with even mainstream media[2] picking up stories of big companies deploying RDF on their Web presence. However, these efforts have mainly concentrated on textual documents in order to annotate concepts like shop opening hours, prices, or contact data. Far fewer occurrences can be noted for RDF video description on the public Web. Related efforts are automatic video content

* This work is partly funded by the EU FP7 I-SEARCH project (project reference 248296)

¶ The author is currently a PhD student at Universitat Politècnica de Catalunya, Department LSI, Campus Nord, Edifici Omega, Jordi Girona 1-3, 08034 Barcelona, Spain.

[1] Live demo at http://tomayac.com/semwebvid/, username: iswc2010, password: iswc2010

[2] http://www.nytimes.com/external/readwriteweb/2010/07/01/01readwriteweb-how-best-buy-is-using-the-semantic-web-23031.html

extraction, or the W3C Ontology for Media Resource, the prior being computationally intensive, and the latter being in non-final status.

In [3] we introduced SemWebVid. The development of SemWebVid was driven by the following objectives:

- Improve **searchability** of video content by extraction of contained entities and disambiguation of those entities (for queries like *videos of Barack Obama where he talks about Afghanistan while being abroad*).
- Enable **graphical representations** of video content through symbolization of entities (for e.g. video archives of keynote speeches where one could *graphically skim through long video sections at a glance*).

These goals can be reached through RDF video descriptions and we thus developed SemWebVid to create such descriptions in a potentially automatable way based on live data found on YouTube.

1.1 SemWebVid Dataflow

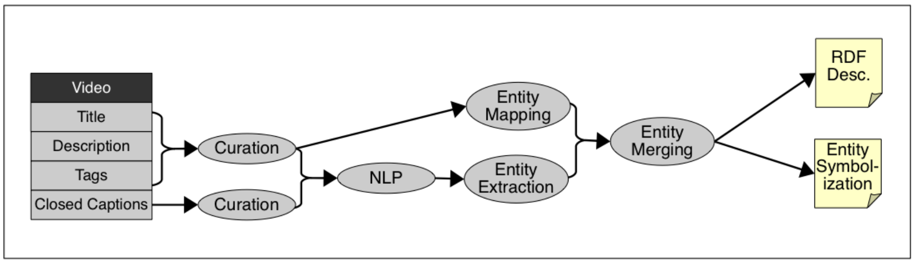


Fig. 1. SemWebVid Dataflow Diagram

The raw data for the RDF video descriptions consist of the beforementioned two pillars: the user-generated video title, video (plaintext) description, and tags on the one hand, and the user- or auto-generated[3] closed captions on the other. The complete dataflow for the application can be see in figure 1.

1.2 Curation of Raw Data

Before the entity mapping and extraction steps, the raw data need to be curated. While the video titles are typically trouble-free as they are usually very descriptive, the main problem with the plaintext video descriptions is that sometimes they get abused for non-related spam-like messages or comments rather than providing a proper summary of the video content. Unfortunately this is hard to detect, so in the end we decided to simply use them as is. With regards to tags the main issue are different tagging styles. As an example see potential tags for the concept of the person Barack Obama:

- `"barack", "obama"` (all words separated, 2 tags)
- `"barack obama"` (space-separated, 1 tag)

[3] YouTube allows for auto-generation of closed captions through speech recognition: http://youtube-global.blogspot.com/2010/03/future-will-be-captioned-improving.html

- "barackobama" (separate words concatenated, 1 tag)

The first split style is especially critical if complete phrase segments are expressed in tag form:

- "one", "small", "step", "for", "a", "man"

For our demonstration we use an API from Bueda[4] in order to split combined tags into their components and try to make sense of split tags. We model tags in the application with the Common Tag vocabulary.

The curation step for closed captions mainly consists of removing speaker and hearable events syntax noise from the plaintext contents, and obviously the cues (time markers for each caption). This can be easily done using regular expressions, the syntax being a variation of ">>Speaker:" and "[Hearable Event]".

1.3 Entity Extraction and Mapping

We try to map the list of curated tags back to entities using plaintext entity mapping Web services[5] from DBpedia [1], Sindice [2], Uberblic, and Freebase. This works quite well for very popular tags (samples below from the DBpedia URI Lookup Web service, all results are prefixed with http://dbpedia.org/resource/"):

- "barack obama" => Barack_Obama

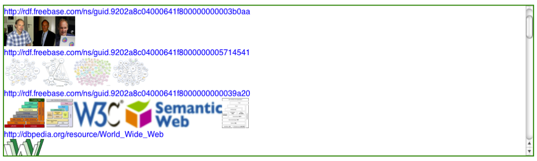
It somewhat succeeds for very generic tags (though with obvious ambiguity issues):

- "obama" => Obama,_Fukui

It fails for specific tags ("ggj09" was a tag for the event "Global Game Jam 2009"):

- "ggj09" => N/A

It is thus very important to preserve **provenance** data in order to judge and estimate the quality of the mapped entities.



http://rdf.freebase.com/ns/guid.9202a8c04000641f80000000003b0aa

http://rdf.freebase.com/ns/guid.9202a8c04000641f8000000005714541

http://rdf.freebase.com/ns/guid.9202a8c04000641f800000000039a20

http://dbpedia.org/resource/World_Wide_Web

Fig. 2. Graphical symbolizations of several entities (TimBL, Semantic Web, etc.)

With regards to the curated closed captions, description, and title we work with NLP Web services[6], namely OpenCalais, Zemanta, and AlchemyAPI. For the test cases that we used for our experiments (famous speeches, keynotes) the detected entities were relatively accurate. As a final step the detected entities from the different NLP Web services are merged, and a symbolization for each entity gets retrieved by means of a heuristic approach, including Google image search. See figure 2 for an example.

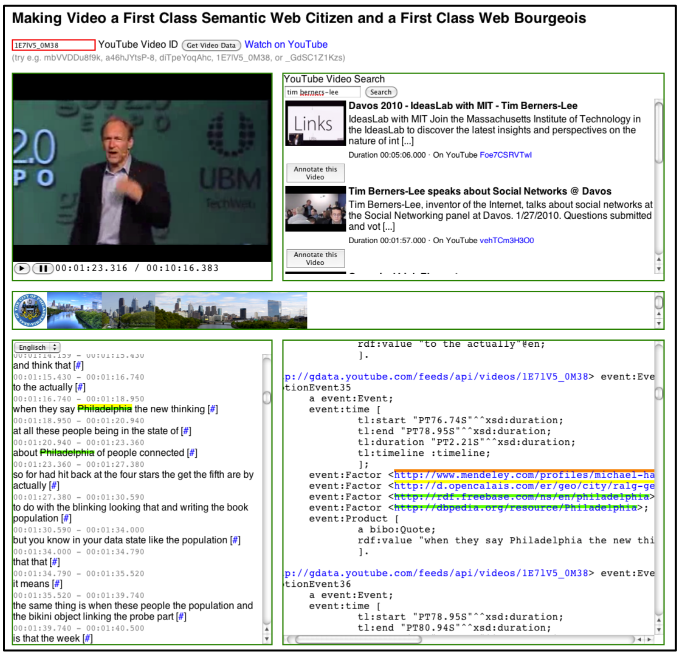
[4] http://www.bueda.com/developers

[5] http://platform.uberblic.org, http://www.freebase.com, http://sindice.com, http://dbpedia.org

[6] http://opencalais.com, http://zemanta.com, http://alchemyapi.com

2 Implementation Details of SemWebVid

SemWebVid is designed to be an online Ajax application for interactive use. Unfortunately the terms and conditions of some of the NLP Web services involved do not allow for a SemWebVid API, however, due to its design both on-the-fly RDF description generation and permanent linking to previously generated descriptions are possible. See figure 3 for a screenshot of the current application.



Making Video a First Class Semantic Web Citizen and a First Class Web Bourgeois

1E7lV5_0M38 YouTube Video ID Watch on YouTube
(try e.g. mbVVDDu8f9k, a46hJYtsP-8, diTpeYoqAhc, 1E7lV5_0M38, or _GdSC1Z1Kzs)

00:01:23.316 / 00:10:16.383

YouTube Video Search
tim berners-lee

Davos 2010 - IdeasLab with MIT - Tim Berners-Lee
IdeasLab with MIT Join the Massachusetts Institute of Technology in the IdeasLab to discover the latest insights and perspectives on the nature of int [...]
Duration 00:05:06.000 · On YouTube Foe7CSRVTwl

Tim Berners-Lee speaks about Social Networks @ Davos
Tim Berners-Lee, inventor of the Internet, talks about social networks at the Social Networking panel at Davos. 1/27/2010. Questions submitted and vot [...]
Duration 00:01:57.000 · On YouTube vehTCm3H3O0

```
00:01:14.159 - 00:01:15.430
and think that [#]
00:01:15.430 - 00:01:16.740
to the actually [#]
00:01:16.740 - 00:01:18.950
when they say Philadelphia the new thinking [#]
00:01:18.950 - 00:01:20.940
at all these people being in the state of [#]
00:01:20.940 - 00:01:23.360
about Philadelphia of people connected [#]
00:01:23.360 - 00:01:27.380
so for had hit back at the four stars the get the fifth are by
actually [#]
00:01:27.380 - 00:01:30.590
to do with the blinking looking that and writing the book
population [#]
00:01:30.590 - 00:01:34.000
but you know in your data state like the population [#]
00:01:34.000 - 00:01:34.790
that that [#]
00:01:34.790 - 00:01:35.520
it means [#]
00:01:35.520 - 00:01:39.740
the same thing is when these people the population and
the bikini object linking the probe part [#]
00:01:39.740 - 00:01:40.500
is that the week [#]
```

```
                rdf:value "to the actually"@en;
                ].

p://gdata.youtube.com/feeds/api/videos/1E7lV5_0M38> event:Eve
tionEvent35
        a event:Event;
        event:time [
                tl:start "PT76.74S"^^xsd:duration;
                tl:end "PT78.95S"^^xsd:duration;
                tl:duration "PT2.21S"^^xsd:duration;
                tl:timeline :timeline;
                ];
        event:Factor <http://www.mendeley.com/profiles/michael-ha
        event:Factor <http://d.opencalais.com/er/geo/city/ralg-ge
        event:Factor <http://rdf.freebase.com/ns/en/philadelphia>
        event:Factor <http://dbpedia.org/resource/Philadelphia>;
        event:Product [
                a bibo:Quote;
                rdf:value "when they say Philadelphia the new thi
                ].

p://gdata.youtube.com/feeds/api/videos/1E7lV5_0M38> event:Eve
tionEvent36
        a event:Event;
        event:time [
                tl:start "PT78.95S"^^xsd:duration;
                tl:end "PT80.94S"^^xsd:duration;
```

Fig. 3. SemWebVid screenshot showing Tim Berners-Lee's infamous potato chips speech at gov2.0 Expo 2010. Below the video box the concept of the city of Philadelphia is symbolized. The left lower box shows the closed captions directly, the right box the RDF description.

2.1 Innovations In the Application

SemWebVid, as mentioned before, is an interactive Ajax application that runs entirely on the client side. This allows for a number of innovations as outlined in the following.

2.1.1 Deep-linkable RDF Video Descriptions

Each RDF video description that gets generated for a video has its separate video-dependent URI, which allows for deep-linking of video descriptions, allowing for passing the state of the application to a different user agent. For example the URI

`http://tomayac.com/semwebvid/#id=3PuHGKnboNY` links to the RDF video description of Barack Obama's inauguration speech.

2.1.2 W3C Media Fragments Working Group Linking Style

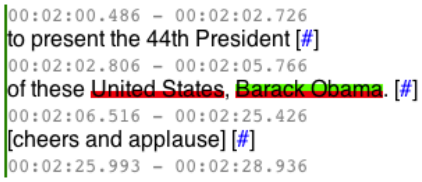
SemWebVid uses the newly proposed means of linking video content introduced by the W3C Media Fragments Working Group. For example

`http://tomayac.com/semwebvid/#id=3PuHGKnboNY;t=10`

links to the scene in the video where Obama starts repeating the inauguration oath phrases, In the long-term this allows for punctually or per scene describing the concrete contents in videos, and having them directly semantically linkable.

2.1.3 Color-coded Provenance

In SemWebVid we introduce a novel way of visually representing provenance data. The problem is that an entity can come from more than one source, for example in the phrase "I, Barack Hussein Obama, do solemnly swear that I will execute the office of President to the United States faithfully", the entity Barack Obama will most probably be recognized by most NLP Web services. We thus need a way to visually represent that. By means of partial highlighting in different colors we can represent up to four sources for one entity. Figure 4 shows the provenance color-coded representation of the entities United States (as matched by Alchemy) and Barack Obama (as matched by Alchemy and Zemanta simultaneously).



00:02:00.486 – 00:02:02.726
to present the 44th President [#]
00:02:02.806 – 00:02:05.766
of these United States, Barack Obama. [#]
00:02:06.516 – 00:02:25.426
[cheers and applause] [#]
00:02:25.993 – 00:02:28.936

Fig. 4. Two entities provenance color-coded

3 Design Choices and Lessons Learnt

We have the current application in use since end of August 2010 and since then had the time to positively judge its overall usefulness, however, could not yet run detailed evaluations of specific points.

3.1 Design Choices

We decided to base our application entirely on JavaScript client side code (except for a trivial PHP proxy server to overcome the Same Origin Policy). This design choice was taken because we wanted the application to be a Web application that could be easily tested by anyone without any installation process whatsoever. In consequence we decided to work with JavaScript's native JSON notation, rather than work with the RDF data provided by some of the Web services involved. This choice was taken simply due to the fact that we only need a simple list of entities with their occurrences in the closed captions, rather than the whole RDF stack. In consequence we do not use any of the JavaScript RDF libraries.

3.2 Lessons Learnt

We found our approach to work well, with necessary improvements in the details listed below.

3.2.1 User Interface Reactivity

The current user interface has the tendency to block during computationally intensive operations like highlighting the detected entities in the closed captions stream using our color-coded provenance technique. For most of our tests we used Google Chrome, however, were successfully able to test the application on Safari and Firefox, each time with the short time blocking user interface. A solution to speed up the user interface is to use Web Workers for computationally intensive tasks, and thereby keep the main user interface reactive.

3.2.2 Entity Mapping Quality

We differentiate between entity mapping based on video tags and entity extraction based on closed captions, video descriptions, and titles. While the quality of entity extraction is good due to the availability of a context, the quality of entity mapping is rather bad due to the complete absence of a context. We are about to evaluate several strategies to improve entity mapping quality, among them majority-based, weight-based winner entity selection processes, and a combination of both. This will be one of the objects of our future research.

4 Conclusion and Future Work

While we are not the first[7] to connect RDF (and thus Linked Data) with video, SemWebVid's contribution is to present an automatic text-based and light-weight way to generate RDF video descriptions. Future work is among other things to determine whether semantic video searchability is superior to existing video search techniques. Concrete next steps include research in order to improve entity mapping quality by means of a combination of several entity mapping Web services on the one hand, and user interface reactivity improvements by moving blocking tasks onto Web Workers. The big vision of SemWebVid is to make it a useful tool for e.g. video archives, we thus need to scale our approach, and evaluate its usefulness on big video archives.

[7] Sack, H: http://www.hpi.uni-potsdam.de/fileadmin/hpi/FG_ITS/papers/Harald/DSMSA09.pdf

Appendix - Open Track Minimal Requirements

- *The application has to be an end-user application, i.e. an application that provides a practical value to general Web users or, if this is not the case, at least to domain experts.*
  Our application is designed for entry-level Semantic Web or multimedia semantics domain experts, however, it is accessible to general Web users as well and due to its interactive nature its usefulness can be seen relatively quickly.
- *The information sources used should be under diverse ownership or control.*
  We use a big variety of information sources reaching from DBpedia to Sindice to Zemanta and OpenCalais to more esoteric services like Bueda. Ownership of the data sources is both academic and commercial.
- *The information sources used should be heterogeneous (syntactically, structurally, and semantically).*
  We work with the diverse formats and natures of the data sources involved. We use the JSON representation of the particular data, and where this is not possible, we work with an XML representation that we convert to JSON.
- *The information sources used should contain substantial quantities of real world data (i.e. not toy examples).*
  We use live video data as found on YouTube. The application contains a search field that allows for direct annotation of the found videos.
- *The meaning of data has to play a central role.*
  Our application is all about giving meaning to video content.
- *Meaning must be represented using Semantic Web technologies.*
  We represent meaning in several RDF mark-up formats, we support RDF/N3, RDF/Turtle, RDF/XML, and RDF/N-Triples. The descriptions can be downloaded.
- *Data must be manipulated/processed in interesting ways to derive useful information.*
  We pre-process the raw data for our application in order to get the most meaningful results. We call this curation. The paper contains the details.
- *The semantic information processing has to play a central role in achieving things that alternative technologies cannot do as well, or at all.*
  Similar results to ours could be achieved using very computationally intensive video content analysis and audio recognition. Our approach is more light-weight and so far provides great results.

Appendix - Open Track Additional Desirable Features

- *The application provides an attractive and functional Web interface (for human users).*
  We are still at an early stage, however, already now the user interface is accessible to everyone, and with its very solid Ajax foundation it can be adapted quickly to new user interfaces.
- *The application should be scalable (in terms of the amount of data used and in terms of distributed components working together).*

At present we are prohibited by our data sources' terms and conditions to scale, as they prohibit us from deploying their service in a different service. We may, however, combine several services under the same roof, which we do with our application.

- *Ideally, the application should use all data that is currently published on the Semantic Web.*
  Our application contributes even more data to the Semantic Web, interlinking the data published so far.
- *Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies, or validate the results obtained.*
  We are in the early stages of evaluating the benefits, however, first results are promising. More intense and at scale evaluations have to take place in future.
- *Novelty, in applying semantic technology to a domain or task that have not been considered before.*
  Applying Semantic Web technologies to multimedia content is a relatively new field, we have provided pointers to related work in the paper.
- *Functionality is different from or goes beyond pure information retrieval.*
  With SemWebVid we give general Web users a tool at hand to produce data for the Semantic Web.
- *The application has clear commercial potential and/or large existing user base.*
  Having cleared the legal requirements with the terms and conditions of the data sources in use, the tool could find its place in e.g. video archives like the BBC's in order to find video content based on semantic technologies.
- *Multimedia documents are used in some way.*
  The whole point of the application is to annotate multimedia documents.
- *The results should be as accurate as possible (e.g. use a ranking of results according to context).*
  We are working towards making the results more accurate, already now we base our judgments in the application on multiple data sources which we still need to rank accordingly, this is mentioned as future research direction.
- *There is support for multiple languages and accessibility on a range of devices.*
  With the current design the application can be easily ported to mobile devices by changing the user interface to a mobile interface and dealing with the lower performance of mobile devices by moving computationally intensive tasks to Web Workers, a change due anyhow.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. In: Proc. of 6th Int. Semantic Web Conf., 2nd Asian Semantic Web Conf., November 2008, pp. 722–735.
2. Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., Tummarello, G.: Sindice.com – A Document-oriented Lookup Index for Open Linked Data. International Journal of Metadata, Semantics and Ontologies, 3 (1), 2008.
3. Steiner, T.: SemWebVid - Making Video a First Class Semantic Web Citizen and a First Class Web Bourgeois, 9th International Semantic Web Conference (ISWC2010), Posters and Demonstrations Track, Shanghai, China, 2010.