Enriching Electronic Programming Guides with Web Data

Pedro Macedo, Jorge Cardoso, Alexandre Miguel Pinto

Dept. Informatics Engineering, University of Coimbra, Coimbra, Portugal,
pmosm@student.dei.uc.pt,{jcardoso, ampinto}@dei.uc.pt

Abstract. Electronic Programming Guide (EPG) systems are software applications running on set-top boxes providing broadcast programming and scheduling information to users of TV services. Although Internet users and enterprises continuously generate massive amounts of web data, a fragment of which is related to TV programs, current EPG systems do not take advantage of this information to enrich the data they provide to users. Addressing this untapped opportunity, we have developed a prototype that gathers web data from several Internet sources, in a variety of formats, and integrates it with the information provided by EPG systems to users. We describe our prototype, evaluate the enrichments achieved, and present our results and directions for future work.
Keywords: Television; Content enrichment; EPG; Linked Data.

1 Introduction

Nowadays, the content that television (TV) stations produce and provide is usually not limited only to the TV business; many such companies provide further information and services over the internet. On the other hand, internet users can produce a significant amount of data related to TV content through collaborative services, like blogs, news/articles records, social networks and on-line ranking systems. Although the generation of information is dynamic and has multiples sources, it is not usually contextualized/linked. This lack of integration makes it is easy for parts of the information to be left unconsumed by users. I.e., two different pieces of information regarding a same subject may not be somehow connected to each other, or even with a same reference to that specific subject. The existence of two islands of information opens the door to a service enrichment opportunity which we take in our work.

The Portuguese TV service provider MEO[1] has an Electronic Programming Guide (EPG)[2]. The lack of contextualization of information is visible in MEO: its EPG information is poor and typically contains only the basic data about each TV program. Linking Internet and EPG content, i.e., performing the enrichment of TV programs' information with Internet material, would allow us to bring added value to TV services for viewers by increasing TV contents contextualization.

[1] http://meo.pt/

[2] Electronic Programming Guides provide users of television with continuously updated menus displaying broadcast programming or scheduling information.

Recently, PT Innovation[3] has been investing in the improvement of its MEO service with new technologies and software. In the process, we have identified this opportunity to enrich MEO with information freely available on the Internet, e.g., from sources such as Linked Data (or Linked Open Data - LOD)[4] [3], in order to provide a better TV experience and service, and as a result we produced the improved EPG+[5].

In this paper we detail the development of the EPG+ system from its definition to implementation. In Section 2 we provide a detailed view of the problem and objectives. A summary of similar projects that have been developed regarding the television and video enrichment are presented in Section 3. Section 4 explains the use case chosen and how the EPG+ system will be integrated with the MEO service. In Section 5 we present the system architecture defined as well as the solution adopted. We detail the technologies used for the implementation in Section 6. Section 7 presents two of the evaluations made to the prototype developed. Section 8 summarizes possible improvements to our approach considered for future work. The Conclusions in Section 9 close this paper.

2 Problem Definition and Objectives

On a daily basis, MEO viewers often use the EPG to select which channel/program they will watch. However, MEO's EPG contains poor information regarding TV programs broadcasts. Figure 1 illustrates how scarce the information provided is. Although

FOX How I Met Your Mother T2 - Ep. 10
Start: 15:33 Duration: 24 min
Ted searches for the woman of his dreams in New York City, accompanied by his four best friends: Marshall, Lilly, Robin and Barney

Fig. 1. Details of a broadcast provided by MEO EPG

the EPG structure is able to support more details, only about half of the data fields are currently being used. Our goal is to enrich the EPG (creating the EPG+) thus allowing MEO clients to make better decisions regardless of whether they would like to watch/record a program or not.

In this project we focus the enrichment process on a smaller part of the problem by enriching only soccer games and TV series. Our enrichment is made based on news articles extracted from pre-defined Internet sources and with semantic annotations to resources from DBPedia[6].

The goals of the work described here in are:

[3] The Research division of Portugal Telecom, `http://www.ptinovacao.pt`
[4] `http://linkeddata.org/`
[5] EPG+ stands for Electronic Program Guide Enriched
[6] Linked Data version of Wikipedia `http://dbpedia.org/About`

1. **Enrichment of the EPG with Internet content.** There are some technical difficulties in the extraction of information from different sources such as API's, web services, RSS[7] feeds or even scrapping. Since each TV program has its own broadcast time, reasoning is needed to determine whether the content extracted is outdated or not.
2. **Enrichment of the EPG with LOD concepts.** We want the system to create links between the content of the EPG and concepts of the LOD cloud. Those links allow a wider knowledge and contextualization of TV programs.

Both objectives share a common problem which is the language of content: MEO is provided to Portuguese citizens and, therefore, its contents are provided in its native language. Thus, the enrichment must be based on Internet contents in Portuguese. This requirement creates barriers regarding the usage of external tools for annotation of content and for processing natural language as well as the sources to be used to collect external information.

We believe that the enrichment process we developed can be expanded to other domains besides TV programs. E.g., other types of multimedia content, such as music tracks, software and its documentation, services, among others.

3 Related Work

One of the main projects regarding the enrichment of video/television content is NoTube [1,5,2]. The main goal of NoTube is to prepare TV for the future Internet addressing challenges of TV content ubiquity and choice, personalization and integration. NoTube enriches its EPG by linking its contents and metadata to concepts from Link Data cloud using LUPedia [8]. In our project we do this type of enrichment in a very similar way. Since LUPedia does not support the portuguese language we had to use other external services with the same functionalities as LUPedia. Another component on NoTube's enrichment is the personalization regarding each user. NoTube collects and processes data from social networks, profiles each user and then suggest contents available on the EPG or in their system. NoTube also have an ANTS [9] (Automatic Newscast Transcription System) system that analyses the video stream, identifies the broadcast of newscast and tags them. In accordance with the profile of each user different news can be delivered, and that is how our solution differs from NoTube's approach. Our goal is to enrich the EPG with news articles related to each TV program and to provide them to every user. Also, the news articles used are collected from web sources and not extracted from the broadcast of a given channel.

Choudhury and Breslin [4] developed a framework to annotate and retrieve web videos with light semantics. The framework collects metadata about the videos through APIs and RSS feeds from different sources , models the data according to an ontology, processes the content for concept extraction, integrates the data with Linked Data cloud and, at last, extracts the existing semantic relation between content and information.

[7] Rich Site Summary, a web feed format used to publish frequently updated works in a standardized format

[8] `http://lupedia.ontotext.com/`

The content is processed for concept learning, where textual term tag approach and entity recognition using Open Calais[9]. I.e., it identifies in a text entities such as person, location, events, among others and associates them with predefined tags of the system. The next step is the integration of data with the Linked Data cloud where they link their content and DBPedia resources.

It is understandable the current need to improve the user experience while he interacts with media video. The scope of this project is different from ours, but the enrichment philosophy is similar: to have video/television content related to on-line information. Even using different external services (DBPediaSpotlight[10] and AlchemyAPI[11]) for the identification and annotation with DBPedia resources our solution does part of the EPG enrichment based on Choudhury and Breslin approach.

4 Approach

Our enrichment system is intended to be integrated with the current EPG system provided by MEO. In Figure 2 we present the integration scenario of the EPG+ system with the current system.

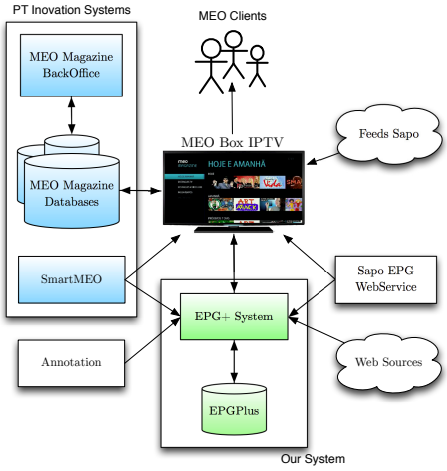

Fig. 2. Integration of EPG+ with PT Innovation systems

MEO clients access the EPG information through the set-top box (STB) which is implemented over the Microsoft Mediaroom Presentation Framework[12]. This is an In-

[9] http://www.opencalais.com/

[10] A tool for automatically annotating mentions of DBpedia resources in text spotlight.dbpedia.org/

[11] http://www.alchemyapi.com/

[12] http://www.microsoft.com/mediaroom/

ternet Protocol Television (IPTV) framework that allows an easy implementation of featured applications, including MEO EPG. There are multiple sources feeding the STB with information; e.g., MEO Magazine which is currently being managed by PT Innovation. Also, it has diverse applications that use some feeds provided by SAPO[13] (e.g. entertainment, economics, highlights, among others). Our EPG+ system follows an integration scheme similar to that of MEO Magazine. It runs in the back-end collecting data, performing the enrichment and storing it in its own database. Then the MEO STB sends requests to our system about information regarding the EPG+.

Since we propose to enrich the current MEO EPG, which is provided by SAPO EPG service, our system also uses the information from SAPO. After collecting the EPG, a classification of each broadcast (e.g. movie, documentary, sports, among others) is done. This classification of programs allows us to know the nature of each broadcast and possibly relate various broadcasts.

The types of programs to be enriched with information available on the internet were restricted to soccer games and TV series. Our enrichment uses only news articles from a pre-determined set of sources for each category to be enriched. The extraction of news articles is continuous and collects various information pieces such as title, subtitle, description, images and video links. This is done via an annotation process on each television program and news articles contents. This annotation process identifies the named entities present in its textual content (e.g. title, description and summaries) and links them with resources available on LOD datasets. To make those annotations we used two services available on-line: DBPediaSpotLight and AlchemyAPI. AlchemyAPI is a free tool that identifies entities, and since we are dealing with text written in Portuguese language, it was a determinant factor. After detecting the entities present in the text provided, we use DPBPediaSpotLight to annotate the entity with links to the wikipedia. To finalize the enrichment process, the association between news articles and TV programs is made through the matching between the annotations made to both contents. A list of ordered news articles related to each TV program is obtained.

5 System Architecture

The system gathers, independently of any other process, new information from SAPO EPG every six hours, classifies it and then stores the results in the database. The extraction of news articles is independent, and for each Internet source we defined an hourly verification for new information, via RSS or HTML scrapping, and collected the new articles. The information is mapped into the EPG+ data structure and then sent to a system buffer which contains all news articles collected. Then, the system controls which television programs are eligible for the enrichment process and searches for news articles, from the buffer, which are related to the TV program itself. When a relation is found, the news article is added to the TV program and then stored into the database.

We now detail in the next sections the system's architecture and the construction of the EPG+ data model.

[13] SAPO is another company of the PT group. `http://www.sapo.pt`

5.1 Architecture

We use the FMC (Fundamental Modeling Concepts) graphical notation [7,8] for the simple definition of the system's architecture we present in figure 3.

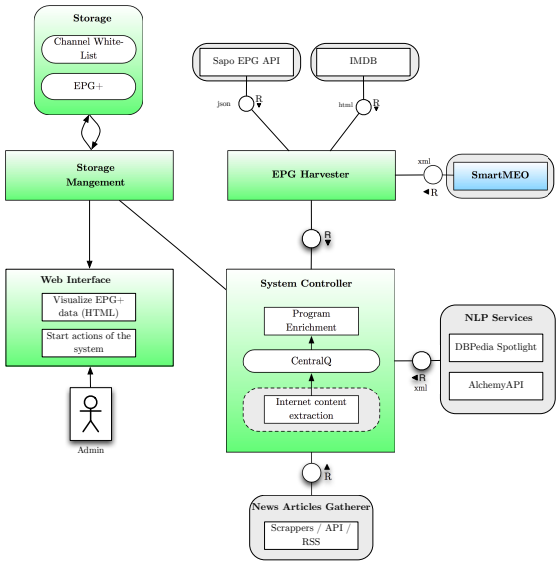

Fig. 3. EPG+ system architecture

This architecture is composed of four main components, of which we detail only two: the *EPG Harvester* and the *System Controller*. The remaining two, Web Interface and Storage Management, are just ordinary components that implement a simple web page interface to browse data and the mechanisms to retrieve and store data into a relational database, respectively.

The *EPG Harvester* gathers EPG information from SAPO EPG through a web service[14]. An internal *EPG Harvester* module extracts, by requesting SAPO EPG web-service, the EPG information regarding each channel and for each TV program and classifies the program using PT Innovation's classifier SmartMEO and using IMDB information to confirm that movies are not misclassified as TV series and vice-versa. After the harvesting and classification the data is sent to the *Storage Management* and then stored in the database.

The *System Controller* is the central broker and controls all the other components. It contains a module that keeps track of the current state of the EPG+ and, when it needs to be updated, every six hours, it makes a request to the *EPG Harvester* component

[14] `https://store.services.sapo.pt/pt/Catalog/other/meo-epg/technical-description`

for new information. It also contains the gathering modules which feed the system with news articles extracted from the Internet. There is an instantiation of the module *News Articles Gatherer* for each category (TV series and soccer games), and they provide the system the information that is used to the enrichment. The information is collected from the external sources through API, RSS or HTML scrappers, and published to a central message queue. The content of each news article is sent to the annotation module which, with the usage of two external services, retrieves the entities present in the content and links them with LOD concepts. The content is attached to a new message and published in the central queue. This component manages the selection of eligible programs to be annotated and enriched.

5.2 Data Model Construction

In order for our data model to be able to store the two types of information (MEO EPG and news articles) needed to enrich the TV programs in MEO's EPG, it has also to be able to represent the current MEO EPG un-enriched. The EPG describes several channels, which transmit hundreds of TV programs. These have basic data already in the EPG, such as *title*, *description* or *broadcast time*. Programs can be transmitted repeatedly in any channel and they can have contextual meaning with other TV programs. E.g., a TV program may belong to a series belonging to a specific season. This information is present in the title of the TV program but only in a textual format, e.g., *How I Met Your Mother S1 Ep. 15*. Since the MEO EPG is based on SAPO EPG, and we have access only to the last one, we conducted our study with SAPO EPG.

Regarding the information extracted from the Internet, it differs from source to source and it can also change over time. E.g., each news article always has a title, but it may not have a subtitle (each source may or may not provide it). The model needs to be flexible and generic enough to accept the usage of any news article source and, at the same time, to be able to represent the information they may provide.

We conducted an analysis for each principal source, and identified and extracted the main attributes. We have taken into account the structure of SAPO EPG as well as the characteristics that each program may have. We also conducted a study on a set of Internet sources which provided content regarding the following categories: soccer, TV series, movies and soap operas. Table 1 shows a comparison between RSS feeds and information from HTML pages.

Figure 4 shows the final EPG+ data model developed for the system. It supports both enriched and non-enriched TV programs, and it is oriented to TV programs, unlike SAPO EPG. It provides a new level of contextualization between the same programs with different broadcasts and a new structure that supports the identification of TV series associated with seasons and episodes.

Table 1. Analysis of information available in HTML and RSS

Internet Sources	HTML available fields										
	Title	Subtitle	Summary	Description	Source	Date	Category	Article Link	Image Link	Video link	Rating
A Bola	•	NA	NA	•	•	•	NA	•	•	•	NA
O Jogo	•	NA	•	•	•	•	•	•	•	•	NA
Sapo Desporto	•	NA	•	•	•	•	•	•	•	•	NA
Mais Futebol	•	•	•	•	•	•	•	•	•	•	NA
Zero Zero	•	NA	NA	•	•	•	•	•	•	NA	•
Sapo Cinemas	•	NA	•	•	•	•	•	•	•	•	•
Ipsilon	•	NA	•	•	•	•	NA	•	•	•	NA
Cinebox	•	NA	•	•	•	•	NA	•	•	•	NA
MSN Cinema	•	NA	•	•	•	•	NA	•	•	•	NA
TV Prime	•	NA	NA	•	•	•	•	•	•	•	NA
Sllex Series	•	NA	•	•	•	•	NA	•	•	•	NA
MSN TV Series	•	NA	•	•	•	•	NA	•	•	•	NA
Sapo Telenovelas	•	NA	•	•	•	•	•	•	•	•	NA
Total	13/13	1/13	10/13	13/13	13/13	13/13	7/13	13/13	13/13	12/13	2/13
RSS format	•	NA	NA	•	•	•	•	•	•	NA	•

Legend: • - contains
NA - Not Applicable

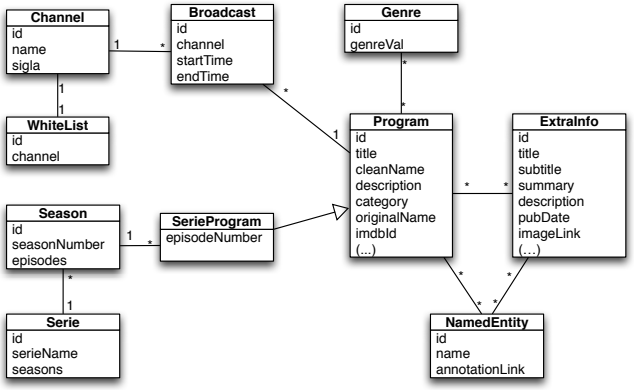

Fig. 4. The EPG+ Data Model

6 Implementation

Given the characteristics of the system we used JBoss AS [15] and ESB [16] since they encapsulate and provide a wide variety of enterprise-level technology services useful for the implementation of our system (e.g. web services, messaging services, persistence, among others). The usage of these proven technologies lends robustness to our system, endowing it with a high level of tolerance to changes and adaptability to new sources or features.

In this section we only detail the implementation of the most important modules of the system which are depicted in Figure 3.

The *EPG Harvester* component is prepared to request, receive and handle the data from SAPO EPG every six hours. To reduce the data redundancy that is requested/received,

[15] `http://www.jboss.org/jbossas`
[16] Enterprise Service Bus

we implemented the Polling Consumer [6] pattern. When the *System Controller* verifies that the data in the EPG+ is outdated, it requests the *EPG Harvester* to start its harvesting with the time interval desired and waits for the response with the new EPG information. Our system is flexible enough to accommodate the possibility of adding or switching the EPG data source. In order to allow such flexibility, we used JAX-Web Services [10] in the construction of our System Controller providing it with an interface to request the update of the EPG for each source. Every interval hour the system starts the processes to gather the most recent news articles from all external sources (soccer and TV series). Any particular source can publish its news articles at any time and the system must gather each article as soon as possible. The solution we implemented to cater for this situation is based on the Publish-Subscriber Channel [6] pattern using a JMS (Java Messaging Service) queue for communication. Each *Internet content extraction* module retrieves the news articles from the source, maps it into a message object and publishes it to the JMS queue *CentralQ*. To complete the pattern, when a TV program is sent to the *Program Enrichment* module in order to be enriched, a consumer is created which reads from the *CentralQ* the news articles and proceeds with the enrichment process. To optimize this process we implemented the Selective Consumer [6] pattern, which only reads the message objects matching a set of given criteria. The *EPG Harvester* and *System Controller* components also implement the Content Enricher pattern which allows the augmentation of data, e.g., when adding a news article to an already existing un-enriched TV program.

7 Evaluation

We performed several types of evaluations of our system: validation of the EPG+ data model, evaluation of the classification and enrichment of TV programs.

7.1 Data Model

The data model evaluation is split into two different processes: 1) functional validation with system usage; and 2) data model support for new external sources.

Functional validation aims at assessing whether the data model is able to support all of the system's functionality. Table 2 presents some counting results regarding the data in EPG+ database at the time of the writing of this document.

	Enriched	Classified	Total	O Jogo	Mais Futebol	Sapo Desporto	TV Prime	Sllex
Programs	141	1537	1697					
Broadcasts			6393					
TV Series			46					
Extra Infos			1026	519	292	102	111	3

Table 2. General statistics of the EPG+ data

With a total of 1697 different TV programs, and over 6300 different broadcasts, this indicates that each TV program is transmitted an average of four times. Since our model

is oriented to TV programs and each one has a list of *broadcasts*, it is possible to reduce information redundancy and at the same time have a new contextualization over time of broadcasts. It also stores 46 TV series – this was possible since the model is able to represent a TV series with its different *seasons* and respective *episodes*. This is a major improvement in the contextualization of each episode and also allows future new enrichments oriented to entire seasons or series. Finally, all the news articles extracted from the predefined sources were successfully stored and the relations to their respective TV programs established. With each program having a list of enrichments (*extraInfo* field), it is possible to maintain a history line of news articles associated to a specific program and, therefore, if the TV program has a new broadcast it is possible to access previous enrichments.

For the validation of the data model support for new external sources, we checked if it was able to represent information from a new category that was not considered during the definition: *Music*. We chose two sources of news articles for this category: MTV[17] and SAPO Músicas[18]. Table 3 shows the information fields each source provides and the comparison with the EPG+ data model.

Table 3. Fields from Music sources and EPG+ data model

	HTML available fields								
Internet Sources	**Title**	**Summary**	**Description**	**Source**	**Date**	**Article Link**	**Image Link**	**Video link**	**Rating**
Sapo Músicas	•	•	•	•	•	•	•	•	NA
MTV	•	•	•	•	•	•	•	•	•
EPG+ Data Model	•	•	•	•	•	•	•	•	NA

Legend: • - contains
NA - Not Applicable

This validation consisted in a detailed analysis of the characteristics of the news articles provided from the two sources chosen to check if the data model could represent all the information. The only field that our data model did not support was the rating. This field was discarded during the construction of the model. Even without running the system to enrich data of TV programs in the EPG using these two new sources, we have good reasons to believe, based on the previous analysis, that our data model will be capable to capture all the information that a new source may provide.

7.2 Enrichment

In order to assess the quality of the enrichment process, we needed to collect enough data. We used a different dataset built over two weeks of early June 2013. The current set has a total of 36 enriched TV programs with a total of 228 different news articles

[17] `http://www.mtv.pt/noticias/`
[18] `http://musica.sapo.pt/noticias`

associated. The analysis process allowed us to verify the true and false positive enrichments, i.e. the enrichments that are correctly related to the TV program and the ones that were wrongly associated by some misleading entity. During this evaluation we verified that nearly 50% of the news articles extracted were associated with a TV program in each enrichment process. With a total of 36 TV programs, 228 news articles, 155 True Positive enrichments, and 73 False Positive enrichments we reach the following conclusions. The results show that approximately 68% (155/228) of enrichments are correct with 155 news articles associated to the respective TV program. Nevertheless, there is a high percentage, approx. 32% (73/228), of false positive enrichments, which led us to analyze their cause. They were originated from the enrichment of soccer games and caused by two determinant factors. The dataset corresponds early June which coincides with the end of the season of almost all soccer competitions. Because of this, the national teams played several games, either as friendly matches or as part of international competitions. Also, it is the period when teams start their player transfers and the majority of the news articles are focused on that subject. These news articles often mentioned the country where the players will be playing for the next season and this is how the connection between the television program and the news articles is made. The second factor is the lack of contextualization that each program of the EPG+ has since there is a national team for each sport type. It is possible to have two different games where the national team of Portugal will be playing but where each game is from a different sport type. Even with this error in the enrichment, PT Innovation stakeholders considered that the 32% of false positives were acceptable since the news articles will be stored and correctly annotated and can be used by future systems.

8 Future Work

Despite the positive results, improvements can be considered for future work. In order to have a better disambiguation between similar TV programs and to improve the enrichment, we believe that a semantic solution should be implemented in order to give an extra contextual knowledge to the EPG+. The exploration of Linked Data, introducing a semantic enrichment, can enhance the EPG+ to a new level of experience and depth of information enabling browsing over related information and background. One important component that could be improved is the gathering of news articles. As stated before, the information retrieval is done by fetching HTML pages which are susceptible to change – we had the opportunity to confirm this when video links were being added to a page. The implementation of a dynamic information gathering from Internet sources would improve the quality of the enrichments. Finally, PT Innovation will integrate this system with its home-TV set-top box based MEO service.

9 Conclusions

We have presented and detailed a possible solution to enrich the information of TV programs in the current EPG of MEO. We do this in two different ways: by gathering data from news articles, and by incorporating information from the web of Linked Data.

This second approach provides an open door to further improvements of the enrichment of the EGP information and its contextualization, namely resorting to Semantic Web technologies. The implementation of our system based on enterprise solutions and technologies allowed us to have a prototype that is able to adapt itself to new sources of data or expand its enrichment to other categories. The next step is to implement the integration of EPG+ system with the current MEO service and then provide the enriched TV service to MEO clients.

10 Acknowledgements

This work is partially supported by iCIS project (CENTRO-07-ST24-FEDER-002003) which is co-financed by QREN, in the scope of the Mais Centro Program and FEDER, and by CISUC, financed by FEDER funds via POFC – COMPETE and by FCT, project FCOMP-01-0124-FEDER-022703. Also a special thanks to Eng. Telma Mota, for providing the possibility to be a part of this project and for all the assistance received from PT Innovation.

References

1. L. Aroyo, L. Nixon, and S. Dietze. Television and the future internet: the NoTube project. In *Future Internet Symposium (FIS) 2009*.
2. L. Aroyo, L. Nixon, and L. Miller. NoTube: The television experience enhanced by online social and semantic data. In *IEEE Intl. Conf. on Consumer Electronics - Berlin*, pages 269–273, 2011.
3. C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *IntL. J. on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, Mar 2009.
4. S. Choudhury and J. G. Breslin. Enriching videos with light semantics. In *Intl. Conf. on Advances in Semantic Processing (SEMAPRO 2010)*.
5. B. Schopman et al. NoTube: making the Web part of personalised TV. 2010.
6. G. Hohpe and B. Woolf. *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2003.
7. A. Knöpfel, B. Gröne, and P. Tabeling. *Fundamental Modeling Concepts. Effective Communication of IT Systems*. John Wiley, 2005.
8. A. Knöpfel. FMC Quick Introduction. `http://www.fmc-modeling.org/quick-intro`. Last visited in 21 January 2013.
9. A. Messina, R. Borgotallo, G. Dimino, L. Boch, and D.A. Gnota. An automatic indexing system for tv newscasts. In *IEEE Intl. Conf. on Multimedia and Expo*, pages 1595 –1596.
10. Oracle. The Java EE 6 Tutorial: Part III Web Services. `http://docs.oracle.com/javaee/6/tutorial/doc/`, 2013. Last visited in 12 June 2013.