Improving Television Program Guides by Accessing Linked Data Resources in OntoTV System

José Luis Redondo-García[1], Alvaro E. Prieto[1], Adolfo Lozano-Tello[1],

[1] Universidad de Extremadura. Quercus Software Engineering Group
Escuela Politécnica, Campus Universitario s/n, 10071
Cáceres, Spain
{jluisred, aeprieto, alozano}@unex.es

Abstract. Nowadays, digital television offers a huge variety of content, and the number of available channels and platforms grows day by day. In this situation, viewers need all the possible information for making decisions about what they want to watch. OntoTV is a television content management system that collects data about programs from various existing sources and represents this information by using knowledge engineering and ontologies.

As the available programs' descriptions are usually scarce, this system incorporates mechanisms to complete them when possible. In this document, a new component for OntoTV system that retrieves additional information about movies will be described. This component takes advantage of the Linked Data's benefits, in order to provide a more complete service to the viewers, which will enjoy much more information about the movies they watch.

Keywords: Television, Linked Data, Program Guide, Ontology, Movie.

1 Introduction

The audiovisual sector, particularly television, occupies a prominent place in the information society. The number of digital TV platforms that offer lots of content to their users is growing every day. However, this wide variety of television contents makes it more difficult for information to be correctly structured and easily accessed. Also, information is usually scattered and incomplete. Imagine that one of the available channels broadcasts the movie "Blade Runner". The user will usually find a very brief description about it. If he wants more information, like other movies also filmed by Ridley Scott, he has to resort to alternative sources like Internet searches.

OntoTV [1] is a system for managing television content information that has been designed to solve these problems. This system integrates various available television data sources, and when information is scarce, it accesses external resources to gather extra details in a timely and transparent fashion. OntoTV uses classic software components called crawlers, which retrieve information from certain sites that use their own way to represent the information. This process usually has a high

computational cost due to the fact that it is necessary to take into account all the particularities about the way every source represents the information.

But today a new way of publishing structured information, called Linked Data [2], is being used. This form of representation proposes some principles for structuring and interlinking data, which becomes more useful. It builds upon standard Web technologies, such as HTTP and URI so information is shared in a way that can be read automatically by computers. This enables data from different sources to be easily connected and queried. It is possible to add new Linked Data components in OntoTV system that implement strategies for retrieving television information, with better results than classic crawlers' ones, and with lower computational cost.

All this collected information has high relevance not only to be showed to the user on the television screen, but also to be stored in a knowledge base where advanced operations can be made. Specifically, in this research, ontologies have been used in order to represent the collected content information. This type of formal representation of knowledge shrinks the semantic gap in search and recommendation operations, making them as effective as possible.

In the areas of multimedia and digital television, there are already some examples of systems that are similar to OntoTV. NoTube [3] is a television content management system that is able to access a great number of data sources: electronic guides for various digital platforms, users' preferences on sites like Twitter and Facebook, and even Linked Data resources [4]. Also, the Sensee system [5] integrates different television resources, such as programming from the BBC, XMLTV guides, movies' descriptions from IMDB website, as well as Linked Data datasets like Geo Ontology and Time Ontology. However, these systems do not elaborate on the way content descriptions are retrieved from television platforms or give detailed description about how to access to Linked Data sources. For these reasons, this research shows how OntoTV system retrieves information about contents and how enriches it by gathering extra information about movies from Linked Data datasets.

2 OntoTV System

As seen in the introduction, there are already some research projects that collect information about television contents from various data sources, including Linked Data ones. Then, all this information is offered to the viewer. But none of these systems gives a high-detailed description of the collection process that has been used for obtaining suitable and valuable information.

OntoTV system (ONTOlogy-based management system for digital TeleVision) is a television content management system that collects information about television programs, and stores it by using ontologies [6]. This way, OntoTV can offer detailed content descriptions to the viewers, as well as execute search and recommendation operations with a high degree of personalization. This system was previously introduced in [1], where the ontology-based knowledge model that represents information about contents and viewers was described. However, for this Linked Data focused research a different television domain ontology, called "BBC Programmes", has been used. This ontology is provided by the BBC (British Broadcasting

Corporation)[1]. The BBC is a public broadcaster that has extensive experience in making its television data available in the web by using semantic technologies, so this ontology has been considered as the most appropriate for representing television programs and channels in a more standard way.

The main objective in this research is the improvement of OntoTV's input data module, in order to add to the knowledge base as information as possible. This module is able to read television information from various digital platforms like Spanish DTT (Digital Terrestrial Television), as well as from some sources on Internet where additional programming guides are offered. Now, new features will be included in this input module so it will be able to retrieve information from Linked Data datasets for enriching movie descriptions.

2.1 Data Collection Module

The "Data Collection" module retrieves information about television contents, which later will be represented in the knowledge base by using ontologies. This module needs a file format for managing the data from the different sources and incorporating it into the knowledge base. XMLTV[2] format has been chosen, because even it is very simple, it is widely used and clearly focused on the television domain.

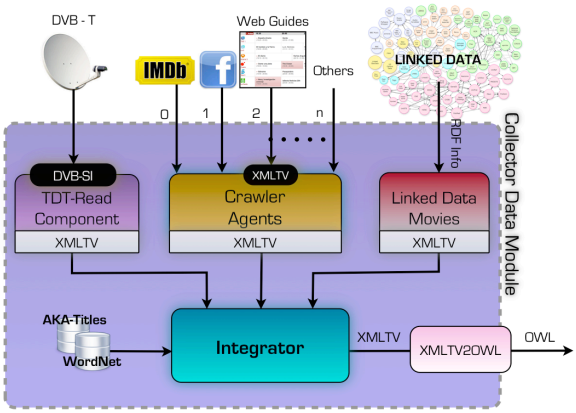

Fig. 1. Schema of the components inside of "Data Collection" module.

A deeper view of Figure 1 shows that there are various components inside of "Collection Data" module. First, "Read-type" components extract data directly from some of the digital TV platforms supported by the system. "CrawlingAgents" include a set of agents that connect to alternative Internet sources by using the TCP/IP protocol, "LinkedData Movies" accesses Linked Data Resources for improving the information about movies, and "Integrator" elaborates a single data file that contains

[1] http://www.bbc.co.uk/ontologies/programmes/2009-09-07.shtml
[2] http://wiki.xmltv.org/index.php/Main_Page

all the information collected by the three previous components. Next section describes the complete process of collecting information, and also the way all these components work together in retrieving valid information about television contents.

3 Collecting Information about Television Contents in OntoTV

"Collection Data" module performs a list of steps in order to retrieve information from the several available sources. This section describes all these steps:

1. *Read a Programming Guide from any of the Available Platforms*. The first step consist of retrieving information about contents from the original sources. For example, OntoTV has specific hardware for receiving Digital Terrestrial Television, which is the public television platform in Spain. OntoTV has a component called "DTT-Read-Component" that implements a mechanism for accessing information contained in DTT stream by using DVB-SI and DVB-EIT, as described in [1]. This information is translated into XMLTV format.
2. *Accessing alternative programming guides from Internet sources*. Unfortunately, the original programming guides retrieved from content providers' sources are scarce. As a result, The "Integrator" component request "CrawlingAgents" to resort to external resources where more detailed content information is available:
   - The web page "La Guía TV" (www.laguiatv.com). It stores schedules and content descriptions of programs for major television channels in Spain. There is a crawler that uses the open source package xmltv-0.5.59 for converting the HTML code of this page into a XMLTV file.
   - The web page "Mi Guía TV" (www.miguiatv.com). Similar to the above case.
   - Windows Media Center. Microsoft offers programming guides for the main TV channels in Spain. OntoTV accesses to this data and translated it into XMLTV.
3. *Merging all the collected programming guides*. The main objetive in this step is to take all the descriptions that belong to the same content and create a single XMLTV programming guide. Figure 2 shows descriptions for Blade Runner:

```
<!-- LAGUIATV.COM -->
<programme start="20101214220000 +0100" channel="CLa
    <title lang="es">El cine de La 2: Blade Runner</
    <category lang="es">pelicula</category>
</programme>
```

```
<!--WINDOWS MEDIA CENTER -->
<programme start="20101214220000 +0100" stop="201012
    <title lang="es">El cine de La 2</title>
    <desc lang="es">Espacio que incluye la emisión de
    <date>20070427</date>
    <category lang="es">Otro</category>
    <category lang="es">Película</category>
    <length units="minutes">120</length>
</programme>
```

```
<!-- DTT READER -->
<programme channel="15" start="20101214210940" stop=
    <title>Blade Runner</title>
    <sub-title></sub-title>
    <desc></desc>
</programme>
```

```
<!-- MIGUIATV.COM -->
 <!--NO INFORMATION RETRIEVED -->
```

Fig. 2. Descriptions for the movie Blade Runner broadcasted on "La 2" at 22:00 pm.

The "Integrator" component determines which descriptions from different sources are refering to the same content, by using certain similarity criteria described in

[7]. Then, all these description for different sources are merged together by selecting the most appropriate attributes [7] from every of the identified data fragments. The result is shown in Figure 3:

```
<!-- FUSION XMLTV -->
<programme start="20101214220000 +0100" stop="20101214235000 +0100" channel="fu
    <title lang="es">El cine de La 2: Blade Runner</title>
    <desc lang="es">Espacio que incluye la emisión de una película.</desc>
    <date>20070427</date>
    <category lang="es">Otro</category>
    <category lang="es">Película</category>
        <category lang="es">pelicula</category>
    <length units="minutes">120</length>
</programme>
```

Fig. 3. Description for the movie Blade Runner after the merging process.

The same is made with every of the contents in the retrieved programming guides. This way, after the merging process, a unique XMLTV document is obtained.

4. *To enrich movie descriptions by accessing Linked Data resources*. The "Integrator" component has a XMLTV file that is more complete than the original ones. However the guide may still have some empty attributes, so additional enriching processes should be performed. In this last step, a component improves descriptions about movies. At this point, it is easy to determine if a content is a film or not by reading the "category" attribute, so "Integrator" component realises that fact and asks "LinkedData Movies" component for more information. This step is described in more detail in the next part of the document.
5. Once the collected data is in a single XMLTV file, this information has to be added into the knowledge base. The XMLTV2OWL component translates the XMLTV "programme" elements into instances of the BBC's OWL Ontology.

4 Enriching Movie Descriptions by Accessing Linked Data Resources

This section describes the way "LinkedData Movies" component accesses Linked Data Resources, identifies information about films, and complete the movie descriptions of the XMLTV programming guide.

Before explain the selected method, it is necessary to say that there are other alternatives that have been finally rejected because they had disadvantages. For example, one of the solutions was to perform search operations in LinkedMDB[3] dataset by executing SPARQL queries on its endpoint. However, this catalog has no entries for various films, so more complete datasets are needed.

Another approach was the method described in the book "Linked Data: Evolving the Web into a Global Data Space" [8], which uses the Crawling Pattern. It consists of creating our own mashup by using two open-source tools, which access the Web of

[3] http://www.linkedmdb.org/

Data and store all the information in local storage for further processing. But this method has a high computational and spatial cost. The crawling process is slow and must be repeated periodically to ensure the information is not outdated. If there is not a clear idea of what data that is going to be retrieved, and many complex queries have to be executed, this is a good alternative. But in this research the application is focused only on film descriptions so these kinds of processes are not so efficient.

5.1 Accessing the Semantic Mashup SIG.MA

As seen in the previous paragraph, to create our own information store is not the best solution. But luckily, there are some global mashups that can be used online. In this research, the semantic mashup called SIG.MA (http://sig.ma/) has been used.

The advantages are clear: it is possible to access relevant information from various datasets without perform a slow crawling process. The "LinkedData Movies" component is released of intensive operation. This approach also recognizes language alternative movie titles, due to the fact that SIG.MA is able to collect information in different languages, and the results usually include references to the original version film's descriptions, by using `http://www.w3.org/2002/07/owl#sameAs` links.

The "LinkedData Movies" component has been coded using the Java language. It carry out two important task that are described in the two following subsections.

4.1.1 Retrieving RDF File from Server

The basic mechanism for accessing Linked Data on the Web is to dereference HTTP URIs into RDF. In the case of SIG.MA mashup, it is necessary to make requests with the URL http://sig.ma/search?q=moviename, where "moviename" is a string that represents the movie's title. The Java code below is able to retrieve RDF files instead of the HTML representation that web browsers usually use:

Java Code that retrieves RDF files from SIG.MA's server.

```
import org.apache.commons.httpclient.*;
HttpClient client = new HttpClient();
HttpMethod mth = new GetMethod(urlsigma + filmName);
mth.addRequestHeader("Accept", "application/rdf+xml");
int responseCode = client.executeMethod(method);
InputStream is = method.getResponseBodyAsStream();
OutputStream os = new FileOutputStream(rdfFile);
//Write RDF to FILE using well-know methods
```

After setting the variable "filmName" to "Blade Runner" and executing this code, a RDF file is obtained. It contains information about this famous movie.

4.1.2 Extracting Information from RDF Files by using SPARQL Queries

The RDF file that describes the movie is already available, so data can be extracted from it by using SPARQL language. As the component has been coded in Java, we

have used the library JENA ARQ[4] that make easier the process of executing queries on RDF files.

Java Code that executes queries on the RDF files to extract the desired information.

```
import com.hp.hpl.jena.query.*;
Model m;
m = ModelFactory.createMemModelMaker().createModel("");
m.read(RDFfile,null);
//Execute Query
Query query = QueryFactory.create(stquery);
QueryExecution qe;
qe = QueryExecutionFactory.create(query, m);
ResultSet results = qe.execSelect();
qe.close();
while (results .hasNext()){ ...}
```

The above code executes the SPARQL query expressed in the "stquery" string. For example, in Figure 4 the query for retrieving information about the director of "Blade Runner" movie and the results of the execution are shown. Now, "Linked Data Movies" component knows that Blade Runner's director is Ridley Scott.

```
PREFIX sigma: http://sig.ma/property/
PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#
SELECT ?director ?name
WHERE {
    ?film  sigma:director ?director.
    ?director rdfs:label ?name.
    }
Retrieving info from SIG.MA...
-------------------------------------------------------------------
| director                                  | name               |
===================================================================
| <http://dbpedia.org/resource/Ridley_Scott> | "Sir Ridley Scott" |
-------------------------------------------------------------------
```

Fig. 4. SPARQL query for retrieving the name of the movie's director.

There are other movie items that can be collected by using the same strategy described before. Table 1 shows these items, next to their corresponding XMLTV elements and SIG.MA vocabulary's properties.

Table 1. Movie items that can be also retrieved from SIG.MA and stored in XMLTV format.

Movie Item	XMLTV Element	SIG.MA Property
Language	tv.programme.language	<sigma:language>
Length	tv.programme.length	<sigma:runtime>
Country	tv.programme.country	<sigma:country
Rating	tv.programme.rating	<sigma:ratings>
Director	tv.programme.credits.director	<sigma:director>
Actor	tv.programme.credits.actor	<sigma:starring>

[4] http://jena.sourceforge.net/ARQ/

Writer	tv.programme.credits.writer	<sigma:writer>
Producer	tv.programme.credits.producer	<sigma:producer>
Composer	tv.programme.credits.composer	<sigma:music_composer>
Image	tv.programme.icon	<sigma:picture>

5.2 Navigating around other Datasets

Linked Data makes it possible to navigate around the global knowledge. For this reason, information that SIG.MA mashup offers can be extended by accessing alternative documents that are referenced in the RDF result file. For example, not only the name of the Blade Runner's director can be retrieved, but also a URI that points to other document where more details about the director are available.

Taking a deeper look at the RDF data, the Ridley Scott's URI references DBpedia's dataset. The mechanism for retrieving data from this source is similar to the previous case, but taking into account that, when building the queries, DBpedia vocabulary has to be used instead of SIG.MA's one. Figure 5 shows a query that gets more information about the movie's director:

```
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?date
WHERE {
    ?director  dbpedia-owl:birthDate ?date.
    }
Finding for more information in http://dbpedia.org/resource/Ridley_Scott
-----------------------------------------------------
| date                                              |
=====================================================
| "1937-11-30"^^<http://www.w3.org/2001/XMLSchema#date> |
-----------------------------------------------------
```

Fig. 5. SPARQL query for retrieving the director's date of birth.

5 Results

Below, the final XMLTV description for Blade Runner movie is shown. The attributes that were empty or inexistent are now filled with data, as seen in Figure 6. It only remains to store this programming guide on the BBC ontology by translating the "programme" XMLTV elements into "Programme" classes in BBC ontology, and the same with "channel" XMLTV elements and "broadcaster" classes. There will be now more information in the knowledge base than at the beginning of the collecting and enriching processes, so more interesting operations could be performed:

```
<!-- FINAL XMLTV -->
<programme start="20101214220000 +0100" stop="20101214235000 +0100" channel="fusion0
    <title lang="es">El cine de La 2: Blade Runner</title>
    <desc lang="es">Espacio que incluye la emisión de una película.</desc>
    <category lang="es">Otro</category> <category lang="es">Película</category>
    <category lang="es">pelicula</category>
    <date>20070427</date>
    <language>English</language>
    <country>United States</country>
    <credits>
      <director>Ridley Scott, 1927-11-30, South Shields.</director>
      <actor>Harrison Ford</actor> <actor>Rutger Hauer</actor>
      <actor>Sean Young</actor> <actor>Edward James Olmos</actor>
      <actor>Daryl Hannah</actor> <actor>M. Emmet Walsh</actor>
      <writer>Philip K. Dick</writer> <producer>Michael Deeley</producer>
      <composer>Vangelis</composer>
    </credits>
    <icon src="http://getmovielink.com/images/covers/BladeRunner.jpg" />
    <length units="minutes">120</length>
</programme>
```

Fig. 6. Information about the movie "Blade Runner" collected by OntoTV.

Figure 7 shows the improvements when presenting the information to the user on the TV screen. Graphic Interfaces has been built by using methods described in [9].

Fig. 7. Information that OntoTV offers to the user before and after the enriching process.

6 Conclusion

In this document, OntoTV's data collection module has been shown, in order to explain the way in which information about television content is retrieved. In particular, this research has focused on the enrichment of movies by using Linked Data resources. Both additional programming guides from the Internet and data published in accordance with Linked Data principles has improved the scarce information sent by providers, which has become more comprehensive. Then, this information has been included into the BBC ontology, so it is available for the system, which can perform more accurate operations.

By using Linked Data principles, information is well structured and links between concepts define semantic relations that are not present in classic HTML hyperlinks. This way, it is easier to retrieve the more suitable data at the right time. Also, the decision of using a mashup like SIG.MA has advantages over crawling strategies and queries on SPARQL endpoints. For example, information is retrieved from various datasets, so it is no limited to only one source, and SIG.MA uses an extensive vocabulary that identifies every search result, avoiding high computational searches.

There are various guidelines that may be developed in the future. The most important one is to improve the integration process that makes possible to add new program instances into the BCC ontology. That process is not mature enough and it is still in a development phase, so it is necessary to continue enhancing content categorization methods, resolving data conflicts when integrating similar programs, etc. Other consideration for the future is to make all the collected information available on the Web according with Linked Data principles. Actually, OntoTV only stores information on the local knowledge base, but if this data is published, it can be accessed by another Linked Data applications. Finally, the movie collection process can be extended for providing similar functionalities with series, sport events, etc.

In conclusion, it has been probed that these semantic technologies are very suitable for collecting information in the television field, as occurs with OntoTV system. By using this new strategies for retrieving data about contents, viewers can enjoy a new way of watching television.

Acknowledgments. This work is supported by the research grant PD10006 from Junta de Extremadura, and has been funded by Junta de Extremadura and FEDER (the European Regional Development Fund), under contract TIN2008-02985.

References

1. Redondo-Garcia, J.L, Valiente-Rocha, P., Lozano-Tello, A.: Ontology-based system for content management in Digital Television. CISTI, pp. 277-–283, (2010).
2. Berners-Lee, T.: Linked Data. International Journal on Semantic Web and Information Systems, vol. 4, no. 2, W3C (2006).
3. Schopman, B., Brickley, D., Aroyo, L., van Aart, C., Buser, V., Siebes, R.: NoTube: making the Web part of personalised TV. Proceedings of the WebSci10 (2010).
4. Buser, V.: NoTube: experimenting with Linked Data to improve user experience, Summer School on Multimedia Semantics, Amsterdam, 3 September 2010
5. Bellekens, P., Aroyo, L., Houben, G., Kaptein, A., van der Sluijs, K.: Semantics-Based Framework for Personalized Access to TV Content. Springer, ISWC, pp. 887-–894, (2007).
6. Gomez-Perez, A., Corcho, O., and Fernandez-Lopez, M.: Ontological Engineering: With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web. Springer-Verlag, New York (2004).
7. Redondo-Garcia, J.L., Lozano-Tello, A.: Recolección de Datos sobre Contenidos Televisivos en el Sistema OntoTV. CISTI, Accepted in press (2011).
8. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool (2011).
9. Redondo-García, J.L., González-Sánchez, J.L., Gazo-Cervero, A., Corral-García, J.: New Digital Television and the Interactivity in its Multimedia Applications. SIGMAP, (2009).