FP7-ICT Strategic Targeted Research Project TrendMiner (No. 287863)
Large-scale, Cross-lingual Trend Mining and Summarisation of Real-time Media Streams

D2.4 – Integration of new lexical and terminological data in TrendMiner

Maciej Ogrodniczuk (IPIPAN), Thierry Declerck (DFKI)

Abstract

FP7-ICT Strategic Targeted Research Project TrendMiner (No. 287863)
Deliverable D2.4 (WP 2)

We describe in deliverable D2.4 how the new partners (Daedalus, UC3M, IPIPAN and RILMTA) integrated their lexical and terminological data in the running TrendMiner project. We opted for this task to follow a strategy based on adding labels to existing or to be created knowledge sources that are relevant to the project. With this strategy we ensure a higher level of multilingualism for the distinct use cases of TrendMiner and give a solid basis for supporting cross-lingual applications.
We describe also the adaptation work that was necessary from some of the existing knowledge sources, like the political ontology deployed in the context of WP7.

Keyword list: multilingualism, lexicons, terminologies, ontologies

Nature: Report
Dissemination: PU
Contractual date of delivery: 31 Oct 2014
Actual date of delivery: 0.6.11.2014
Reviewed By: RILMTA, UC3M

TrendMiner Consortium

This document is part of the TrendMiner research project (No. 287863), partially funded by the FP7-ICT Programme.

DFKI GmbH
Language Technology Lab
Stuhlsatzenhausweg 3
D-66123 Saarbrcken
Germany
Contact person: Thierry Declerck
E-mail: declerck@dfki.de

University of Southampton
Southampton SO17 1BJ
UK
Contact person: Mahesan Niranjan
E-mail: mn@ecs.soton.ac.uk

Internet Memory Research
45 ter rue de la Révolution
F-93100 Montreuil
France
Contact person: France Lafarges
E-mail: contact@internetmemory.org

Eurokleis S.R.L.
Via Giorgio Baglivi, 3
Roma RM
00161 Italia
Contact person: Francesco Bellini
E-mail: info@eurokleis.com

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
UK
Tel: +44 114 222 1930
Fax: +44 114 222 1810
Contact person: Kalina Bontcheva
E-mail: K.Bontcheva@dcs.shef.ac.uk

Ontotext AD
Polygraphia Office Center fl.4,
47A Tsarigradsko Shosse,
Sofia 1504, Bulgaria
Contact person: Atanas Kiryakov
E-mail: naso@sirma.bg

Sora Ogris and Hofinger GmbH
Bennogasse 8/2/16
1080 Wien Austria
Contact person: Christoph Hofinger
E-mail: ch@sora.at

Hardik Fintrade Pvt Ltd.
227, Shree Ram Cloth Market,
Opposite Manilal Mansion,
Revdi Bazar, Ahmedabad 380002
India
Contact person: Suresh Aswani
E-mail: m.aswani@hardikgroup.com

Universidad Carlos III de Madrid
Advanced Databases Group
Department of Computer Science
Avda. de la Universidad, 30
Leganés
Spain
Tel: +34 91 624 95 00
Contact person: Paloma Martínez
E-mail: paloma.martinez@uc3m.es

DAEDALUS S.A.
Avda. De la Albufera, 321, 1st floor
Madrid, E28031
Spain
Contact person: José Luis Martínez
E-mail: jmartinez@daedalus.es

Institute of Computer Science,
Polish Academy of Sciences
Jana Kazimierza 5
01-248 Warszawa
Contact person: Maciej Ogrodniczuk
E-mail: maciej.ogrodniczuk@ipipan.waw.pl

Research Institute for Linguistics of the
Hungarian Academy of Sciences
Benczúr u. 33., H-1068 Budapest, Hungary
Contact person: Tamás Váradi
Email: varadi.tamas@nytud.mta.hu

Executive Summary

This deliverable documents how lexical and terminological data resulting from the contributions of the new partners (UC3M, Daedalus, IPIPAN, and RILMTA) have been integrated in the TrendMiner project.
In a first part we describe how we generated a multilingual ontology for the ATC (Anatomical Therapeutic Chemical**)** codes that are central to the new eHealth use case introduced by the partners UC3M and Daedalus. The details of this use case are given in deliverable D10.1 "Newly generated domain-specific language data and tools".

Another part of the integration consisted in the extension of existing TrendMiner ontologies, more specifically for the political ontology that has been developed for WP7. These extended ontologies feature descriptions specific to new use cases: political (Hungarian and Polish) as well as lexical parts used in processing domain-related content (such as nicknames of politicians frequently used in social media communication). Labels of existing classes have been extended with terms from the new languages (Polish, Hungarian and Spanish), and new classes have been directly equipped with multilingual labels.

A last aspect of the integration work concerned the newly developed lexical resources: polarity lexicons or sentiment dictionaries, used together with ontologies by the language processing chains providing data for use case visualisations.

Contents

1. Introduction

This deliverable documents how lexical and terminological data resulting from the contributions of the new partners (UC3M, Daedalus, IPIPAN, and RILMTA) have been integrated in the TrendMiner project.

In a first part we describe how we generated a multilingual ontology for the ATC (Anatomical Therapeutic Chemical**)** codes that are central to the new eHealth use case introduced by the partners UC3M and Daedalus. The details of this use case are given in deliverable D10.1 "Newly generated domain-specific language data and tools".

Such a multilingual ontology, including in its labels all the new languages of the project, is increasing the level of multilingualism offered by the project and supports cross-lingual applications. As for now no such multilingual ontology or even thesaurus for the ATC codes was available. We consider this generated resource therefore as a very promising result of the (extended) TrendMiner project.

Another part of the integration consisted in the extension of existing TrendMiner ontologies, more specifically for the political ontology that has been developed for WP7. These extended ontologies feature descriptions specific to new use cases: political (Hungarian and Polish) as well as lexical parts used in processing domain-related content (such as nicknames of politicians frequently used in social media communication). Labels of existing classes have been extended with terms from the new languages (Polish, Hungarian and Spanish), and new classes have been directly equipped with multilingual labels. We describe in this deliverable in certain details the work done for the Polish case, and the work done for the Hungarian case was very similar.

A last aspect of the integration work concerned the newly developed lexical resources: polarity lexicons or sentiment dictionaries, used together with ontologies by the language processing chains providing data for use case visualizations. The integration of such lexical data was done based on using the MARL model, which was integrated in the TrendMiner federated ontologies (see the Opinion ontology included in the set of TrendMiner ontologies at http://www.dfki.de/lt/onto/, which is also described in deliverable D2.1.2 "Knowledge and Provenance modelling and Stream Reasoning v2).

2. The multilingual ATC Ontology

The new partners UC3M and Daedalus introduced a new use case to TrendMiner, dealing with the topic of mining the mention of drugs and drug adverse effects in social media. This use case has been described in details in deliverable D10.1 "Newly generated domain-specific language data and tools", and here we focus on the generation of knowledge sources and associated multilingual terminologies that have been performed in other to extend in the future the original application scenario to other languages than Spanish, which was the original language of application of the use case.

As the use case of UC3M and Daedalus was relying on gazetteers that made reference to the ATC classification[1], we decided to investigate if this classification could be made available to the project partners in the form of a multilingual ontology, which can be be in the longer term integrated in the TrendMiner federated ontologies[2].

Fortunately there was already an existing OWL model proposed for (a former version of) ATC: http://www.ebi.ac.uk/Rebholz-srv/atc/. We took this model as a departure point and worked then mainly on obtaining multilingual terms from various sources, including the Wikipedia categories lists of ATC terms in Spanish, Polish and Hungarian. After some implementation done in order to harmonize the format of those lists, a script was written in order to add the gained terms to the labels of the ATC ontology, which was existing only in English. Figure 1 below displays an example of the obtained ontology, where the reader can see how the multilingual terminology is encoded.

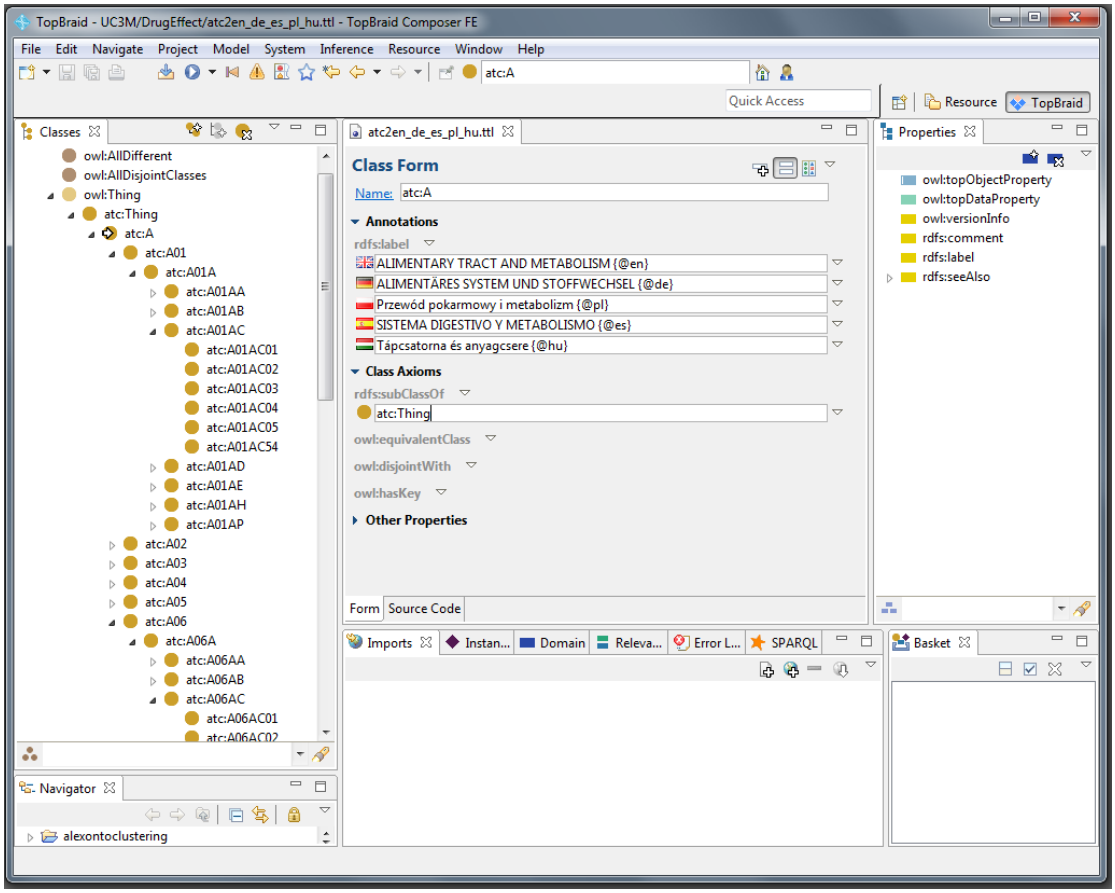

Figure 1: A screen shot from the multilingual ATC ontology, showing how the rdfs:label property is employed for encoding the multilingual terminology

The coverage of languages is still unequal, as we are not working with the same types of sources. For example for German, we could access an updated list of terms delivered by a specialized institute[3], and for Polish and Hungarian, we downloaded first a subset of ATC codes from Wikipedia.

[1] ATC stands for Anatomical Therapeutic Chemical Classification. See also http://www.whocc.no/atc_ddd_index/ for more details.
[2] See http://www.dfki.de/lt/onto/ and deliverable D2.12 for more details.
[3] http://www.dimdi.de/dynamic/de/klassi/downloadcenter/atcddd/version2014/

Before publishing the data, we are consulting first the original author the ATC.OWL model[4]. Our aim is to deliver a high quality multilingual ATC ontology. This is even more important as we discovered that TrendMiner is probably the first project in delivering this kind of multilingual knowledge resource in the context of ATC.

Other (Spanish) lexical resources developed in the context of the eHealth use case are described in D10.1 "Newly generated domain-specific language data and tools".

3. Extensions to the political Ontology

The partners IPIPAN (Poland) and RILMTA (Hungary) contributed to various extents to the WP7 use case, dealing with political topics. Therefore we aimed at extending the language coverage of the existing ontology, which was designed in collaboration between the W2 partner DFKI and WP7 partner SORA. But very soon, it appeared that the first task consisted in extending the ontology in order to cover the country specific political constructs, while a multilingual extension of the generic concepts and shared classes was not an issue.

The Polish extension to the political ontology

The Polish ontology is consistent with the mainstream approach to the research of social structure in social sciences, especially in regard to elite studies. First, it is assumed that significant actors on Polish public (political) scene are all those that have been institutionally based in Polish political and other public institutions – most importantly: legislative, executive and judiciary. These are members of government, deputies and juries. Second, it is assumed that the public actors form their own communication networks within the public sphere. These actors are involved in the public discourse. The data regarding most recognized journalists, reporters and other opinion leaders is to be gathered. The third and last premise is that the public actors may be identified by their relative significance, which is based on being reported by media or public institutions involvement in the decision making processes in Poland. In that regard, the clue representatives of think tanks, NGO's, and private companies were identified. We build the ontology by starting from the institutional point and finish at the decision making part. In this we follow the principles that have been guiding the generation of the first (Austrian) political ontology developed in TrendMiner.

Approach to modelling

In was decided in the Polish TrendMiner team that we want to model not only the current political scene but also its history. Decision was based on the assumption that in the analysis e.g. of political tweets the history can have significant impact. For example, it matters whether the author of a tweet about the government's policy is a former prime minister or a former member of the governing/opposition party. This decision and the specificity of polish political scene have led to changes to the original TrendMiner political ontology that are described below.

[4] Samuel Croset from the European Bioinformatics Institute Cambridge, UK. See also the Github page: https://github.com/loopasam

New data properties	
Property	**Description**
current	domain: Election, range: boolean (to indicate which election which formed the current parliament),
dateFrom and dateTo	domain: PoliticalFunction, range: dateTime (to indicate time span for which a given instance exists)
endReason	domain: PoliticalFunction range: string (reason for which a person stopped to have a given political function)
isLowerHouse	domain: MemberOfParliament range: boolean (to indicate whether an in- stance designates a member of lower or higher house of parliament)
New object properties	
Property	**Description**
hadMember	domain: Group or Organization, range: Person or PoliticalFunction (to indicate past membership)
hasFunction	domain: Person, range: PoliticalFunction (to indicate that a person has/had a given political function)
Modified object properties	
Property	**Description**
hasMember	range was changed from Person to Person or PoliticalFunction
nickname	domain was changed from Person to NamedEntity
occursWith	domain was changed from Political Actor to Political Actor or Political Function
New annotations	
Annotation	**Description**
dateFromAnn	to indicate time span for hadMember property
dateToAnn	
New classes	
Class	**Description**
CurrentMemberOfLowerHouse	subclass of Politician restricted to individuals how are members of lower house in current parliament
CurrentMemberOfHigherHouse	subclass of Politician restricted to individuals how are members of higher house in current parliament
PartyInLowerHouse	subclass of Party restricted to parties whose members are members of lower house in current parliament

PartyInHigherHouse	subclass of Party restricted to parties whose members are members of higher house in current parliament

Table 1. The Polish Political Ontology: changes to the original TrendMiner ontology

Polish political ontology in numbers

The generated ontology contains information about five elections to the Polish parliament that took place since 1997. The earlier elections where omitted due to significant incompleteness of data in the input database. We have gathered information about 1866 people who held function of a member of parliament. 2949 such positions are described. A total of 41 parties existed in the recent history of this political scene, eight of which exist in the current lower house of parliament. Five out of those eight have their members in the higher house of parliament. This includes a special case of non-attached members of parliament.

Sample queries

Parties that have members in current lower house of parliament:

```
party
and hasOrganization
      some ('party faction'
            and hasMember some ('member of parliament'
                                and isLowerHouse value true
                                and occursWith some
                                (Election and current value true)))
```

Members of current higher house of parliament, that are Platforma Obywatelska party:

```
hasFunction some (inverse hasMember
      some (inverse hasOrganization value PO
            and occursWith some (Election and current value true)
            and hasMember some (isLowerHouse value false)))
```

Data sources

1. The Sejm database – with all Polish MPs from 1991 until now was achieved from Sejm (the lower chamber of the Polish parliament). It contained complete information about all Sejm MPs and several cadencies of Senate, so it was semi-manually supplemented with remaining records. The main source of the update was (Kaleta, 2010).
2. Money.pl portal – contains information about Polish parliamentarians from every cadency: http://www.money.pl/gospodarka/politycy/. Web pages was crawled and data parsed by scripts. (see /trendminer/polish_parliament/money_pl).
3. Members of EU Parliament – since the Sejm database did not contain information about Members of EU Parliament, this information was taken from Wikipedia:
   - http://pl.wikipedia.org/wiki/Polscy_pos%C5%82owie_do_Parlamentu_Europejskiego_2004-2009
   - http://pl.wikipedia.org/wiki/Polscy_pos%C5%82owie_do_Parlamentu_Europejskiego_2009-2014

- http://pl.wikipedia.org/wiki/Polscy_pos%C5%82owie_do_Parlamentu_Europejskiego_2014-2019 (see /trendminer/polish_parliament/eu).

4. Various electronic newspaper sources (gazeta.pl, onet.pl, interia.pl, polskatimes.pl, plotek.pl, salon24.pl, tvn24.pl) used to collect nicknames of politicians.

The Hungarian extension to the political ontology

The base for this work was the revised political ontology as generated by the Polish colleagues.

Additions reflect the model of our database of Hungarian politicians and parties during the 2010 and 2014 Hungarian parliamentary elections and the 2014 European Parliament elections. For the 2014 elections, both election nominees and seat holders are represented.

All classes and properties were manually assigned a natural language label in Hungarian (`rdfs:label "..."@hu`).

The following number of individuals was automatically added from RILMTA's database of parties and politicians related to the 2010 and 2014 Hungarian and EP elections:

- 18 instances of Party
- 661 instances of Politician
- 899 instances of Nomination

New classes	
Class	**Description**
MemberOfHungarianParliament	subclass of MemberOfParliament; to distinguish Hungarian MPs from other national MPs
MemberOfHungarianParliament2010	subclass of MemberOfHungarianParliament; to represent MPs that were the members of the Hungarian Parliament in the 2010-2014 term
MemberOfHungarianParliament2014	subclass of MemberOfHungarianParliament; to represent MPs that were the members of the Hungarian Parliament in the 2010-2014 term
MemberOfEuropeanParliament	subclass of MemberOfParliament; to distinguish EP members from national parliament members

MemberOfEuropeanParliament2014	subclass of MemberOfEuropeanParliament; to represent MPs that were the members of the European Parliament in the 2014-2019 term
Nomination	subclass of PoliticalControl; class of instances to represent nomination events that are connector objects for nominating parties (*nominatedBy***)**, nominated politicians (*hasNomination*) and elections nominated at (*nominatedAt*)
ElectionHungarianParliament	subclass of Election; to distinguish Hungarian parliamental elections from European and other national parliamentary elections
ElectionEuropeanParliament	subclass of Election; to distinguish European parliamental elections from other national parliamentary elections
New object properties	
Property	**Description**
hasNomination	domain: Person, range: Nomination; Property to link Nomination instances to Person instances, so Person, Party and Election can be joined together: Person x hasNomination n and n nominatedBy p and n nominatedAt e means "Person x was nominated at Election e by party p"
nominatedBy	domain: Nomination, range Party; Relation between a Nomination instance n and a Party p: the person x that hasNomination n was nominated by p
nominatedAt	domain: Nomination, range: Election; Relation between a Nomination instance i and an Election insance e: the Person that hasNomination i was nominated in election e
New data properties	
Property	**Description**

facebookPage	domain: NamedEntity, range: xsd:string; URL of a facebook page associated to NamedEntity (e.g. Party or Politician)
abbreviatedName	domain: NamedEntity, range: xsd:string; Official abbreviated name of the entity (e.g. party name abbreviations)
nameVariant	domain: NamedEntity, range: xsd:string; A name variant for the entit
New individuals	
Property	**Description**
Hungary	member of Country; created as a value for the hasLocation object property for e.g. political parties (instances of Party) in Hungary
Election_Hungarian_Parliament_2010	member of ElectionHungarianParliament; instance representing the 2010 Hungarian Parliamentary Elections
Election_Hungarian_Parliament_2014	member of ElectionHungarianParliament; instance representing the 2010 Hungarian Parliamentary Elections
Election_European_Parliament_Hungary_2014	member of ElectionEuropeanParliament; instance representing the 2014 European Parliament Elections in Hungary

Table 2. Hungarian Political Ontology: changes to the existing political ontology

Populating the Ontology: an Example

Example: Benedek Jávor was a candidate for a seat in the Hungarian Parliament in 2010 nominated by LMP. In 2014 he ran for a seat in the European Parliament nominated by EGYÜTT-PM. He won both seats.

1. Add LMP and EGYÜTT-PM as instances of Party, and add data properties for name, abbreviatedName, facebookPage(s), plus an object property hasLocation with Hungary:

```
###  http://www.dfki.de/lt/onto/political.owl#LMP

:LMP rdf:type :Party ,
              owl:NamedIndividual ;
     |
     :faceboookPage "https://www.facebook.com/pages/LMP/170605242969861" ,
                    "https://www.facebook.com/lehetmas" ;

     :name "Politics can be different"@en ;

     :abbreviatedName "LMP"@hu ;

     :faceboookPage "https://www.facebook.com/pages/Lehet-M%C3%A1s-a-Politika/101887333185816" ;

     :name "Lehet Más a Politika"@hu ;

     :hasLocation :Hungary .
```

Note: an entity can have more than 1 facebookPage.

```
###  http://www.dfki.de/lt/onto/political.owl#EGYUTT-PM

:EGYUTT-PM rdf:type :Party ,
                    owl:NamedIndividual ;

           :name "Együtt 2014 Mozgalom"@hu ;

           :faceboookPage "https://www.facebook.com/egyuttkorszakvaltok" ;

           :abbreviatedName "EGYÜTT-PM"@hu ;

           :hasLocation :Hungary .
```

2. Add Benedek Jávor as a new instance of Politician and add data properties firstName, lastName, facebookPage. He is also an instance of both MemberOfHungarianParliament2014 and MemberOfEuropeanParliament2014 because he held both seats:

```
###  http://www.dfki.de/lt/onto/political.owl#Javor_Benedek

:Javor_Benedek rdf:type :Politician ,
                        :MemberOfEuropeanParliament2014 ,
                        :MemberOfHungarianParliament2010 ,
                        owl:NamedIndividual ;

               :lastName "Jávor" ;

               :firstName "Benedek" ;

               :faceboookPage "https://www.facebook.com/javorbenedek" .
```

3. Create a new Nomination instance representing his 2010 nomination, using nominatedAt and nominatedBy to mark the election campaign and the nominating party:

```
###  http://www.dfki.de/lt/onto/political.owl#Nomination_Javor_Benedek_hu_2010

:Nomination_Javor_Benedek_hu_2010 rdf:type :Nomination ,
                                           owl:NamedIndividual ;

                                  :nominatedAt :Election_Hungarian_Parliament_2010 ;

                                  :nominatedBy :LMP .
```

4. Create a new Nomination instance representing his 2014 nomination like above:

```
###  http://www.dfki.de/lt/onto/political.owl#Nomination_Javor_Benedek_eu_2014

:Nomination_Javor_Benedek_eu_2014 rdf:type :Nomination ,
                                           owl:NamedIndividual ;

                                  :nominatedBy :EGYUTT-PM ;

                                  :nominatedAt :Election_European_Parliament_Hungary_2014 .
```

5. Add hasNomination data properties for Benedek_Javor linking him to his Nomination instances:

```
###  http://www.dfki.de/lt/onto/political.owl#Javor_Benedek

:Javor_Benedek rdf:type :Politician ,
                        :MemberOfEuropeanParliament2014 ,
                        :MemberOfHungarianParliament2010 ,
                        owl:NamedIndividual ;

               :lastName "Jávor" ;

               :firstName "Benedek" ;

               :faceboookPage "https://www.facebook.com/javorbenedek" ;

               :hasNomination :Nomination_Javor_Benedek_eu_2014 ,
                              :Nomination_Javor_Benedek_hu_2010 .
```

We have been describing in some details the steps that were necessary in order to obtain a joint ontology, but which is still giving place to the description of country specific elements. Concerning the language coverage: as for the ATC ontology, use have been made of the rdfs:label property in order to enter the terminology associated to each class. A script has been implemented in order to add the labels in all the language implied (German, Polish and Hungarian) in the case of shared classes, ensuring thus multilingualism at this level. In the case of the country specific classes, only monolingual labels are available for the time being, but this will be updated. We foresee the publication of our integrated and multilingual political ontology on the TrendMiner page for the 1st of December 2014. As for the multilingual ATC ontology, we are not aware of any multilingual knowledge source that is describing various political systems in the way TrendMiner has been developed it. The 2 screenshots below in Figure 2 and Figure 3show the aspects of the extended political ontology, where the multilingual extensions within the rdfs:label properties can be observed. The reader can also see how the temporal information has been incorporated. We are now populating the ontology with the new members of the Austrian Parliament and will then go for the final integration of all sub-ontologies (the political system of Austria is not included in the version shown in this deliverable, also for reason of place. In fact we still didn't decided if we should have one file with the country-specific ontological elements, or rather to have one file with the common elements and point from there to file with the owl/rdf description of the country specific political systems.

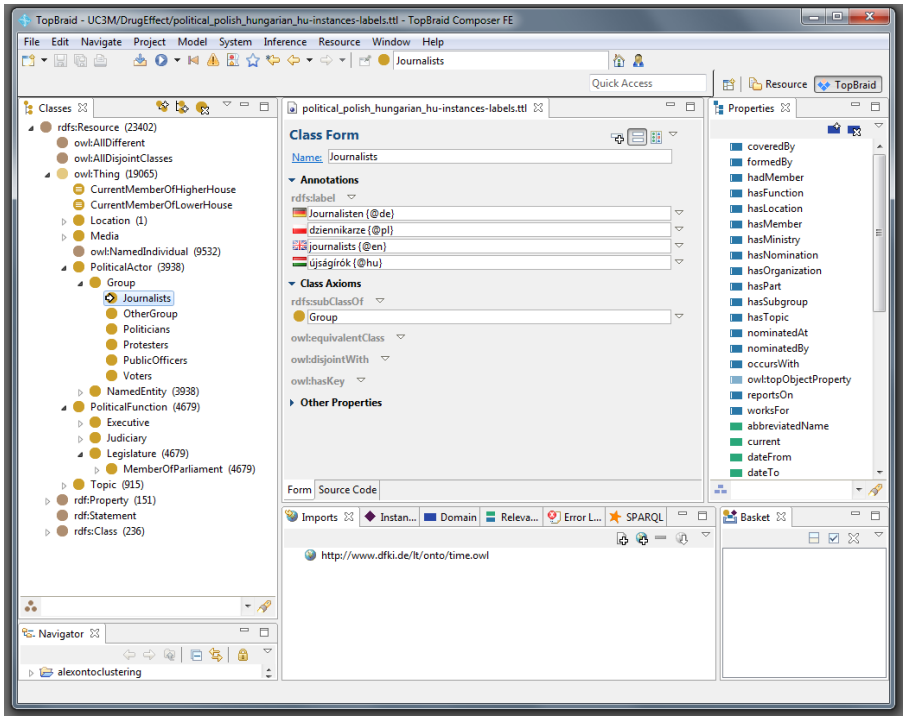

Figure 2: A screen shot of the political ontology extended by the partners IPIPAN and RILMTA. All the classes have been equipped with multilingual labels.

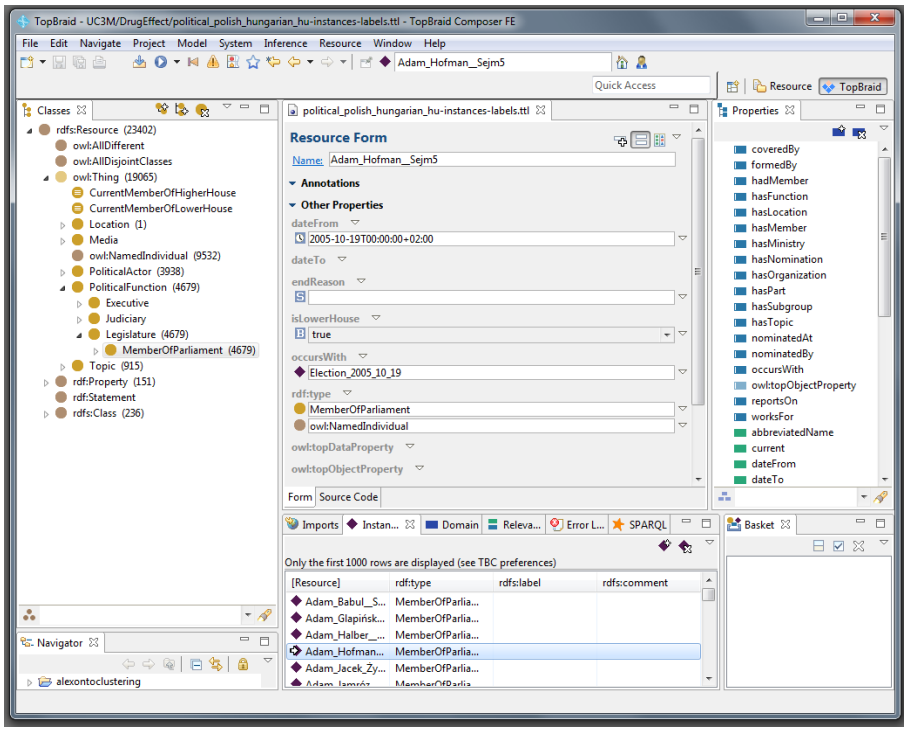

Figure 3: Screen shot of an instance of a Polish MemberOfParliament. The reader can see how we integrate here the temporal information.

4. Polarity Lexicons

Chapter 5 of the deliverable D2.3.2 "Multilingual resources and evaluation of knowledge modelling - v2" describes how a generic polarity lexicon German has been generated on the basis of various previously existing lexicons.

As a next step we modelled the entries of this generic lexicon in order to be compliant to the increasing Linguistic Linked Open Data[5] cloud, using for this the Ontolex codification, which is the result of a joint work of a W3C community group[6]. And as we already integrated the MARL model[7] in the TrendMiner federated ontologies for supported the description of opinions on relevant entities in the TrendMiner use cases, the next logical step seemed to us to combine the use of the Ontolex formal representation of lexical entries and the MARL model. In this we entered in a close cooperation with the (now finalized) project Eurosentiment[8], which is going for a very similar strategy.
Our work consisted in providing a representation of the lexical data using as much as possible information that is available in external resources, like the ISOcat registry[9], and our ontological model for the representation of polarity data. Below, we just display an excerpt of the description of the entry "fehler" (*error*):

```
:LexicalSense_Fehler
      rdf:type lemon:LexicalSense ;
      rdfs:label "fehler"@de ;
      lemon:reference <http://wiktionary.dbpedia.org/page/Fehler-German-Noun-1de> .

:Opinion_Fehler
      rdf:type skosxl:Label , :lemma ;
      rdfs:label "Fehler"@de ;
      hasOpinionObject :Opinion_Fehler_2 , :Opinion_Fehler_3 , :Opinion_Fehler_4 ,
:Opinion_Fehler_1 ;
      :hasPoS <http://www.isocat.org/rest/dc/1333> ;
      skosxl:literalForm "fehler"@de .

:Opinion_Fehler_1
      rdf:type :Opinion_Object ;
      rdfs:label "Fehler"^^xsd:string ;
      op:assessedBy <http://tutorial-topbraid.com/lex_tm#german.lex> ;
      op:hasPolarity op:Negative ;
      op:maxPolarityValue "1.0"^^xsd:double ;
      op:minPolarityValue "-1.0"^^xsd:double ;
      op:polarityValue "-0.7"^^xsd:double .


  ……
```

Figure 4: The RDF, SKOS-XL and lemon representation of the entry, with a link to an ontological framework representing polarity information. The various polarities given by the various sources are represented as "OpinionObjects". The prefix "op" is representing the polarity ontology of TrendMiner

[5] See http://linguistics.okfn.org/resources/llod/ for more details.
[6] See http://www.w3.org/community/ontolex/ for more details.
[7] See http://www.gi2mo.org/marl/0.1/ns.html for more details.
[8] See http://eurosentiment.eu/
[9] See for example http://www.isocat.org/rest/dc/1333 for our selected ISOcat entry for the pos "noun".

Very recent work was then dedicated in linking Polish and Hungarian polarity lexicons available in the TrendMiner consortium to this representation, and to make use of the Ontolex "TranslationVariants" property for marking multilingual equivalents to the German entries encoded so far in our first version of the combination of Ontolex and the opinion ontology. As such our approach is thus again ensuring a close linking of multilingual lexical/terminological data by means of formal representation languages in an ontological context.

The final version of our multilingual polarity lexicon will be published on the TrendMiner page on the 1st of December 2014, after a last review depending on the final release of the Ontolex specifications.

5. Relevance to TrendMiner

Data described in D2.4 are crucial for most scientific and technological objectives of the project, i.e.:

1. Delivered ontologies and lexical data were integrated in TrendMiner, i.e. the political and medical use cases providing means for discovery and tracking of events, trends and attitudes, their development over time, in various media (social networks, blogs, electronic media) and different project languages (English, German, Hungarian, Polish, Spanish).

2. Delivered data (e.g. political ontology) have been multilinguality aligned which will lower the adaptation costs of existing ontology-based information extraction methods.

3. Developed resources (such as instances of the Polish Political Ontology or medical data provided by UC3M) vere actively used for temporal analysis of media streams, offering valuable models of correlations between the text and the changes over time, e.g. political attitudes, popularity polls or tracking drug effects in social media.

4. TrendMiner multilingual ontologies turned out to be basis of the real-time data collection, analysis, summarisation, and search infrastructure over stream media in multiple languages.

5. All methods have been developed by involving key stakeholders from project focus groups. In case of political science, the Political Ontology has been created by representatives of the sociopolitical research community, taking into account the formal constitutional order (with three branches of political power: legislative, executive, judiciary and supplementary institutions) which results in completeness of the formal system.

Relation to other workpackages

D2.4 provides content related to the following TrendMiner workpackages:

- **D5.3.2: Real-Time Stream Media Processing Platform and Cloud Based Deployment – V2** provides implementation of the cloud-based stream media

processing platform which integrates lexical and terminological data described in D2.4

- **D5.5: Deployment of Web services for new cases** provides information about Web services making use of ontologies and lexicons described in D2.4

- **D6.3: Application final results** provides a Web-based prototype for multilingual trend mining and summarisation for financial decision support which makes use of the ontological data described in D2.4

- **D7.3: Application final results** provides the key web-application for gathering multilingual political trends and summaries which was constructed using political ontology and sentiment data described in D2.4

- **D9.1: Integration of annotation generated by annotation tools** contains detailed information of linguistic processing chains using lexical and terminological data described in D2.4

- **D10.1: Harmonization of new languages** contains further information on resources, NLP tools, corpora and evaluation of project ontologies and terminological data cross- and multilingually harmonized within D2.4.

Bibliography and references

Acedański, S. (2010). *A morphosyntactic Brill tagger for inflectional languages*. In Advances in Natural Language Processing, Proceedings of 7th International Conference on NLP, IceTAL 2010, Reykjavik. Lecture Notes in Computer Science, vol. 6233, pp. 3–14. Springer Berlin Heidelberg

Buczyński A. and Wawer A. (2008). *Automated classification of product review sentiments in Polish*. In Kłopotek M.A., Przepiórkowski A., Wierzchoń S.T., and Trojanowski K., (eds.), Intelligent Information Systems, pp. 211–215, Warsaw. Akademicka Oficyna Wydawnicza EXIT.

Hare J. S, Samangooei S., and Dupplaw D. P. (2011). *Openimaj and im- ageterrier: Java libraries and tools for scalable multimedia analysis and indexing of images*. In Proceedings of the 19th ACM international conference on Multimedia, MM '11, pages 691–694, New York, NY, USA. ACM.

Kaleta P. (2010). *Ludzie Władzy Polski Niepodległej 1989-2009*, Zabrze.

Kobyliński, Ł. (2013). *Improving the accuracy of Polish POS tagging by using voting ensembles*. In Vetulani, Z. (ed.), Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, pp. 453–456, Poznań, Poland. Wydawnictwo Poznańskie, Fundacja Uniwersytetu im. Adama Mickiewicza.

Kobyliński Ł. (2014). *PoliTa: A multitagger for Polish*. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, pp. 2949–2954, Reykjavík, Iceland. ELRA.

Kopeć M. (2014). *Zero subject detection for Polish*. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers, pp. 221–225, Gothenburg, Sweden. Association for Computational Linguistics.

Krieger H.-U., Declerck T. (2014). *TMO – The Federated Ontology of the TrendMiner Project*. Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), p. 4164–4171. European Language Resources Association.

Lui M. and Baldwin T. (2012). *Langid.py: An on-the-shelf language identification tool*. In Proceedings of the ACL 2012 System Demonstrations, ACL '12, pp. 25–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Miłkowski M. (2010). *Developing an open-source, rule-based proofreading tool*. Software: Practice and Experience, 40(7):543–566.

O'Connor B., Krieger M., and Ahn D. (2010). *TweetMotif: Exploratory Search and Topic Summarization for Twitt*er. In William W. Cohen, Samuel Gosling, William W. Cohen, and Samuel Gosling (eds.), ICWSM. The AAAI Press.

Ogrodniczuk, M. and Lenart, M. (2013). *A multi-purpose online toolset for NLP applications*. In Métais, E., Meziane, F., Saraee, M., Sugumaran, V., and Vadera, S. (eds.), Proceedings of the 18th International Conference on Applications of Natural Language to Information Systems, volume 7934 of Lecture Notes in Computer Science, pp. 392–395. Springer-Verlag, Berlin, Heidelberg.

Porter M. (1980). *An algorithm for suffix stripping*. Program, 14(3):130–137.

Preotiuc-Pietro D., Samangooei S., Cohn T., Gibbins N., and Niranjan M. (2012). *Trendminer: An architecture for real time analysis of social media text*.

Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B. (eds.) (2011). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Przepiórkowski A., and Buczyński A. (2007). *Spejd: Shallow Parsing and Disambiguation Engine*. In Zygmunt Vetulani (ed.), Proceedings of the 3rd Language & Technology Conference, pp. 340–344, Poznań, Poland.

Radziszewski, A. and Acedański, S. (2012). *Taggers gonna tag: an argument against evaluating disambiguation capacities of morphosyntactic taggers*. In Proceedings of TSD 2012, LNCS. Springer-Verlag.

Radziszewski, A. and Śniatowski, T. (2011). *A Memory-Based Tagger for Polish*. In Proceedings of the LTC 2011. Tagger available at http://nlp.pwr.wroc.pl/redmine/projects/wmbt/wiki/.

Radziszewski, A. (2013). *A tiered CRF tagger for Polish*. In H. Rybiński, M. Kryszkiewicz, M. Niezgódka, R. Bembenik, and Ł. Skonieczny (eds.), Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions. Springer Verlag.

Segura-Bedmar I, Peña-González S, Martínez P (2014): Extracting drug indications and adverse drug reactions from Spanish health social media. In Proceedings of BioNLP 2014, 98-106.

Śniatowski, T. and Piasecki, M. (2012). *Combining Polish morphosyntactic taggers*. In Bouvry, P., Kłopotek, M., Leprévost, F., Marciniak, M., Mykowiecka, A., and Rybiński, H. (eds.), Security and Intelligent Information Systems, volume 7053 of Lecture Notes in Computer Science, pp. 359–369. Springer, Berlin-Heidelberg.

Waszczuk, J. (2012). *Harnessing the CRF complexity with domain-specific constraints*. The case of morphosyntactic tagging of a highly inflected language. In Proceedings of the 24th International Conference on Computational Linguistics (COLING2012), pp. 2789–2804, Mumbai, India.

Waszczuk J., Głowińska K., Savary A., Przepiórkowski A., and Lenart M. (2013). *Annotation tools for syntax and named entities in the National Corpus of Polish*. International Journal of Data Mining, Modelling and Management, 5(2):103–122.

Wawer A. (2012a). *Mining Co-Occurrence Matrices for SO-PMI Paradigm Word Candidates.* In Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL'12 SRW, pp. 74–80, Avignon, France. Association for Computational Linguistics. 3, 44.

Wawer A. (2012b). *Extracting Emotive Patterns for Languages with Rich Morphology*. International Journal of Computational Linguistics and Applications, 3(1).

Wawer A. and Rogozińska D. (2012). *How Much Supervision? Corpus-based Lexeme Sentiment Estimation*. In Data Mining Workshops, 2012 IEEE 12th International Conference on. SENTIRE 2012, ICDMW, pp. 724–730, Los Alamitos, CA, USA, Dec. 2012. IEEE Computer Society.

Woliński, M. (2006). *Morfeusz — a practical tool for the morphological analysis of Polish.* In Kłopotek, M.A., Wierzchoń, S.T., and Trojanowski, K., (eds.), Intelligent Information Processing and Web Mining, Advances in Soft Computing, pp. 503–512. Springer-Verlag, Berlin.