ARHINET

A System for Generating and Processing Semantically-Enhanced Archival eContent

Ioan Salomie, Mihaela Dinsoreanu, Cristina Pop
Sorin Suciu, Tudor Vlad and Ioana Iacob
Department of Computer Science, Technical University of Cluj-Napoca, 15 C. Daicoviciu Street, Cluj-Napoca, Romania

Keywords: Archival domain model, Knowledge acquisition, Historical domain ontology, Semantic annotation, Semantic query, Reasoning.

Abstract: This paper addresses the problem of generating and processing of eContent from archives and digital libraries. We present a system that adds semantic mark-up to the content of historical documents, thus enabling document and knowledge retrieval as response to semantic queries. The system functionality follows two main workflows: eContent generation and knowledge acquisition on one hand and knowledge processing and retrieval on the other hand. Within the first workflow, the relevant domain information is extracted from documents written in natural languages, followed by semantic annotation and domain ontology population. In the second workflow, ontologically guided queries trigger reasoning processes that provide relevant search results.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Most of the existing archives-related information is written in natural language and is available in large amounts, distributed in archives and digital libraries. Although content management systems capable of dealing with distributed data exist for years, making them able to process natural language documents is still a challenge. Documents in natural language human readable format contain unstructured and heterogeneous information, which makes it hard for machines to automatically process their contents.

Our system addresses these challenges by adopting Semantic Web techniques in the context of archive management. We add a layer of machine-processing semantics over the content of raw documents contained in the archives by using (semi) automated knowledge acquisition. The machine-processable semantics, captured in a domain knowledge base, is further used as support for intelligent searches that provide the most relevant results to agent queries. These relevant results represent documents, information and knowledge related to the semantics of the keywords specified in the queries.

In this context, our system offers a solution for managing domain-specific archival data. Our system was specialized to generate archival eContent, to semantically enhance and process it, from the available medieval documents regarding the history of Transylvania. For the semantic enhancement of the documents we have built an ontology core of concepts and relations, which is continuously expanded as new documents are processed. Historians and archivists can use the system to find relevant documents, information and knowledge.

The rest of the paper is organized as follows. In Section 2, we introduce related work. Section 3 presents our generic model of the archival domain. The proposed system architecture is presented in Section 4, while the associated workflows are described in Section 5 and Section 6. Section 7 contains a case study that illustrates the system's functionality in the context of historical archives. The paper ends with our conclusions and future work proposals.

2 RELATED WORK

The OntoPop methodology provides a single-step

Salomie I., Dinsoreanu M., Pop C., Suciu S., Vlad T. and Iacob I.
ARHINET - A System for Generating and Processing Semantically-Enhanced Archival eContent.
DOI: 10.5220/0001833401510158
In *Proceedings of the Fifth International Conference on Web Information Systems and Technologies (WEBIST 2009)*, page
ISBN: 978-989-8111-81-4
Copyright © 2009 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

solution for *(i)* semantically annotating the content of documents and *(ii)* populating the ontology with the new instances found in the documents (Amardeilh, 2007). The solution uses domain-specific knowledge acquisition rules which link the results obtained from the information extraction tools to the ontology elements, thus creating a more formal representation (RDF or OWL) of the document content (Amardeilh, 2006). The OntoPop methodology has certain limitations regarding the resolving of synonyms on one hand and the resolving of multiple instances with the same lexical representation on the other hand. In this paper, we address the identified limitations by extending the OntoPop methodology with new processing steps before populating the ontology.

SOBA is a system designed to create a soccer specific knowledge base from heterogeneous sources (Buitelaar et al., 2006). The system performs *(i)* automatic document retrieval from the Web, *(ii)* linguistic annotation and information extraction using the Heart-of-Gold approach (Schäfer, 2007) and *(iii)* mapping of the annotated document parts on ontology elements (Buitelaar et al., 2006). Our approach performs information extraction from unstructured text and document annotation for a specific domain and uses reasoning on the ontology to infer properties for the newly added instances.

Ontea performs semi-automatic annotation using regular expressions combined with lemmatization and indexing mechanisms (Laclavik et al., 2007). The methodology was implemented and tested on English and Slovak content. Our system was designed to process multilingual documents, including Latin languages, and so far it has provided good results for a corpus of Romanian documents, by using resources specific to the Romanian language.

3 ARCHIVAL DOMAIN MODEL

This paper proposes a generic representation of the archival domain as illustrated in Figure 1. The archival domain is modelled starting from the raw medieval documents provided by the Cluj County National Archives (CCNA, 2008). These documents are hand written and contain many embellishments, making them hard to be automatically processed. Due to this difficulty, in our case studies we have used document summaries generated by the archivists (see Figure 2).

Within our model, the central element is the *document*. Documents belong to a specific domain such as the historical domain or the medical domain. In our research we have used the historical archival domain, formally represented as *domain knowledge* by means of *domain ontology* (*concepts* and *relations)* and *rules*. Documents can be obtained from several *data sources* like external databases, Web sites or digitized manuscripts.

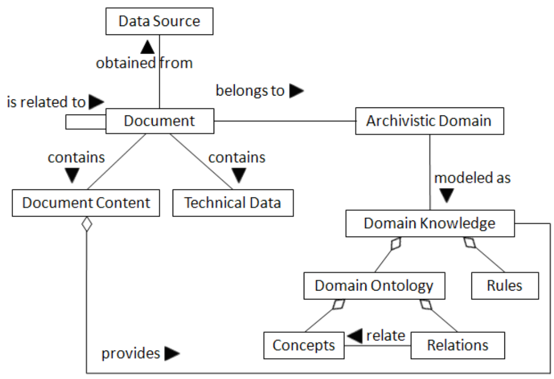
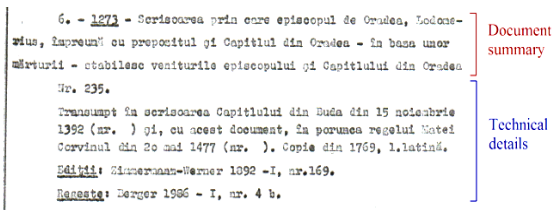


Figure 1: The archival domain model.

The *document content* (see Figure 2) is expressed in natural language in an unstructured manner. In our case study, the document content actually represents a summary of the associated original document. Several documents may be related to one another by referring information about the same topics even if they are not containing the same lexical representations (e.g. names, events, etc.). The document also features a set of *technical data*, such as the date of issue, archival fund or catalogue number. In the case of the document shown in Figure 2, the technical data specifies the document number ("235"), the language in which the raw document was written ("Latin") and the edition in which the original document has appeared ("Zimmermaan-Werner 1892 –I, nr.169").

When searching in the archival documents it is important to identify all documents that are related to a specified topic. To enable information retrieval from all relevant documents, the domain knowledge is used to add a semantic mark-up level to the documents content.


6. - 1273 - Scrisoarea prin care episcopul de Oradea, Lodomerius, împreună cu prepozitul şi Capitlul din Oradea - în baza unor mărturii - stabilesc veniturile episcopului şi Capitlului din Oradea

Nr. 235.

Transumpt în scrisoarea Capitlului din Buda din 15 noiembrie 1392 (nr.) şi, cu acest document, în porunca regelui Matei Corvinul din 20 mai 1477 (nr.). Copie din 1769, l.latină.

Ediţii: Zimmermaan-Werner 1892 -I, nr.169.

Regeste: Berger 1986 - I, nr. 4 b.

Figure 2: Example of a document which contains technical data and the summary of the original archival document.

The domain knowledge core (domain ontology and rules) is captured by processing and analyzing a large repository of archival documents, focusing on identifying their common concepts and relationships. Next, based on information extraction techniques applied on the raw documents, the domain knowledge is enriched through instance population.

4 SYSTEM ARCHITECTURE AND FUNCTIONALITY

The system is structured on three interacting processing layers: the *raw data acquisition and representation layer*, the *knowledge acquisition layer* and the *knowledge processing and retrieval layer*. The layers and their associated resources and processes are shown in Figure 3. The Primary DataBase (PDB) is used for raw document persistence, while the Knowledge Server (KS) is used for learning and reasoning tasks.

The Raw Data Acquisition and Representation layer provides support for collecting and storing data in the Primary DataBase from multiple sources by means of Optical Character Recognition (OCR) techniques on raw documents, data import from external databases or by means of the system's integrated user interface.

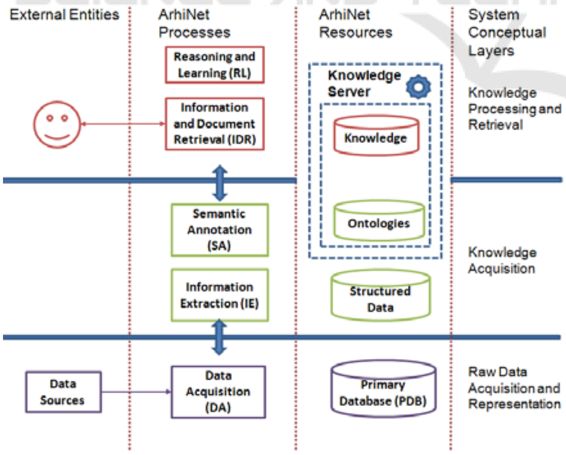


Figure 3: An overview of resources and processes.

The Knowledge Acquisition layer uses pattern-matching to extract relevant data from the raw documents. Based on the domain ontology and on a set of semantic rules, the documents are then semantically annotated. New concepts and instances are identified and added to the domain ontology as a result of this process.

The Knowledge Processing and Retrieval layer enables ontologically-guided intelligent searches over the annotated documents.

The system's main workflows capture the processing steps of the Knowledge Acquisition layer and of the Knowledge Processing and Retrieval layer.

5 KNOWLEDGE ACQUISITION

The objective of the Knowledge Acquisition layer is to extend the domain knowledge by identifying, extracting and annotating the relevant domain-specific information from the summaries of archival documents (see Figure 4).

Knowledge Acquisition uses text mining techniques (tokenization, pattern matching and data structuring processes) applied in a pipeline fashion over the documents' content. Actually the Knowledge Acquisition layer extends the OntoPop (Amardeilh, 2006) (Amardeilh, 2007) methodology with two additional processing steps. The first step, synonyms population, is required for identifying and processing ontology instances having several lexical forms with the same meaning (i.e. they are synonyms) in different documents. For example, the names "Palostelek" present in one document and "Paulusteleky" in another document have been identified and further processed as synonyms. The second step, homonym identification and representation, deals with common lexical representations for different instances. As an example, the name "Mihai" may refer either to the same person or to different persons in different documents.

In the following, we describe the main activities of the knowledge acquisition workflow.

Technical Data Extraction. This activity is responsible for separating the document technical data from its content (for technical data examples see Section 3 and Figure 2).

Lexical Annotation. The objective of this activity is to identify and annotate the relevant lexical elements in the content based on pattern-matching rules. A pattern matching rule defines relationship between the lexical elements and their annotation elements. The output of the lexical annotation activity consists of annotated lexical data represented in a hierarchic format of extracted words along with their annotation elements according to the pattern matching rule (see Figure 8 in the Case Study

SCIENCE AND TECHNOLOGY PUBLICATIONS

Section for an example).

Knowledge Extraction. The objective of knowledge extraction activity is to use the domain ontology in order to semantically annotate the hierarchical structure of annotated lexical elements obtained in the previous activity. This activity is supported by a set of mapping rules. Each mapping rule defines *(i)* ways of associating the annotated lexical elements to ontology concepts and *(ii)* a set of actions for populating the ontology with instances and relations. The result of the knowledge extraction activity is an RDF structure stored in a file associated to the original document content (see Figures 10 and 11 in the Case Study Section for examples).

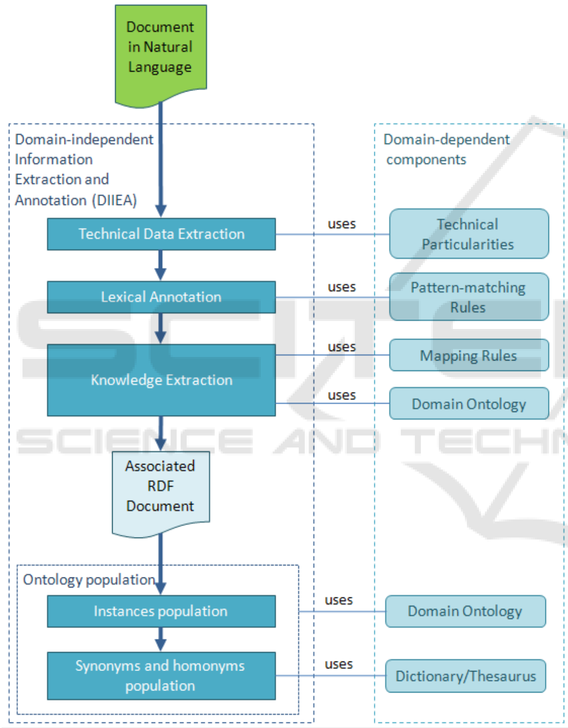


Figure 4: Knowledge acquisition.

Ontology Population and Management. Ontology population activity integrates the new instances, synonyms, homonyms and properties identified in the *knowledge extraction* activity into the domain ontology. By using a dictionary of synonyms, the domain ontology is populated with all the synonyms of an instance. The OWL-Lite ontology representation allows synonym definition through the "sameAs" property, which specifies that an instance X is equivalent with another instance Y. To address the problem of homonyms we defined a distance function that takes as arguments the document technical data, attributes and relations of the ontology-stored potential instances and the current instance being verified for the homonymous relationship. If the computed function value exceeds a certain threshold we consider that the two instances are identical, otherwise they are homonyms.

```
Person(?s) ^ Person(?f) ^ hasSon(?f, ?s) -> hasFather(?s, ?f)

Person(?p) ^ Person(?b) ^ hasBrother(?p, ?b)
                                   -> hasBrother(?b, ?p)
```

Figure 5: Example of logical inference rules.

Ontology management activities aim *(i)* to infer new relations and properties as a result of ontology modification due to previous population processes and *(ii)* to preserve ontology consistency. For the inferring of new properties we have used the Jess rule engine (Sandia National Laboratories, 2008). For example, in a newly processed document we have identified the new instance "Mihail" and its associated property "hasFather" having the range "Albert de Juc". After populating the ontology with this information, the Jess rule engine infers the inverse property "hasSon" with the domain "Albert de Juc" and the range "Mihail". To enable these logical inferences, the Jess rule engine requires SWRL (Horrocks et al., 2004) rules to be defined on the domain ontology. An example of ontology associated SWRL rules is illustrated in Figure 5. The first rule defines the "hasSon" relation between two persons as the inverse of the "hasFather" relation between the same persons. The second rule shows that the "hasBrother" relation between two persons is symmetrical.

6 KNOWLEDGE PROCESSING AND RETRIEVAL

The aim of the *knowledge processing and retrieval layer* is to provide support for intelligent queries that enable searching for the most relevant information available in archival documents. Document searching is performed at two levels: one level relies on the technical data, which narrows the set of documents, while the other level relies on the semantic meaning of the user input query. The user input query triggers a complex reasoning process that includes synonym search, logical inferences and subclass / super class searches. As a result, the set of query relevant documents is identified and new query relevant knowledge may be generated. We used the Jess rule engine and a set of SQWRL (SQWRL, 2008) rules as the main tools for knowledge processing and retrieval.

Usually, in historical documents, several terms, such as person or location names, have different representations around a common root and it is essential to identify all documents containing these synonyms. To address this problem, the initial search query is improved to search for instances connected by the "sameAs" property, thus enhancing the search process. Figure 6 presents an example of a SQWRL enhanced query that enables synonym search. The query searches in the ontology a person that is Magister ("magistru"), taking in consideration all the possible representations (synonyms) of the person's name.

```
Person(?p0) ^ rdfs:label(?p0, ?p0_name)
    ^ sameAs(?p0,?p0s) ^ rdfs:label(?p0s, ?p0s_name)
    ^ hasTitle(?p0,?ha0) ^ rdfs:label(?ha0,"magistru")
        -> sqwrl:selectDistinct(?p0s_name, "Person")
```

Figure 6: Search query example.

The initial search query is further improved by considering a sub-tree of the ontology that contains the searched concept (class). For example, if the instance "Transylvania" belonging to the ontology concept "Principality" is the search key, the query is refined to enable searching for the instances of its super class "TerritorialDivision".

7 CASE STUDY

The system proposed in this paper was used for developing and processing semantically enhanced archival eContent from documents capturing the history of Transylvania, starting from the medieval period. The historical documents have been obtained from the Cluj County National Archives (CCNA, 2008). The original documents are found in Latin, Hungarian, German and Romanian. Each document is associated with a document summary in Romanian which highlights the events and participants. These summaries were used in our system as the *raw documents*, the main source of information.

An example of such a document summary which will be used for further exemplification throughout this section is presented in Figure 7. In English, this summary reads: "Carol Robert, the king of Hungary, donates to Mihail and Nicolae, the sons of Albert of Juk, the Palostelek domain and the Imbuz (Omboz) forest, in the Dabaca County, for their faithful military services carried out together with the magister Stefan, against Moise, a rebel against the crown".

Carol Robert, regele Ungariei, doneaza lui Mihail si Nicolae, fiii lui Albert de Juk mosia Palostelek si padurea Imbuz (Omboz), din comitatul Dabaca, pentru serviciile de arme credincioase aduse alaturi de magistrul Stefan, impotriva lui Moise, razvratit impotriva coroanei.

Figure 7: Medieval document summary.

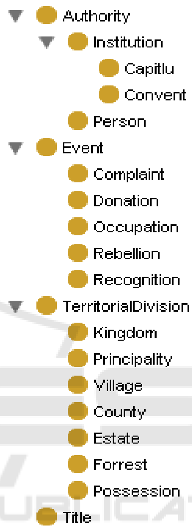


Figure 8: The core of the domain ontology.

In order to specialize the generic knowledge acquisition workflow presented in Figure 4 for the historical domain, we have used a corpus of several documents for creating *(i)* a core for the historical domain ontology, *(ii)* a set of specific pattern-matching rules for the annotation of the lexical elements and *(iii)* a set of mapping rules between the annotated lexical elements and ontological concepts.

Figure 8 presents the core of the domain ontology that was manually created by using the Protégé Ontology Editor (Horridge et al., 2007), based on the analysis of several raw documents.
We created the set of specific pattern-matching rules for the annotation of lexical elements as *JAPE grammars* (Tablan et al., 2004). A JAPE grammar groups in *phases* the *rules* that specify actions to be performed when certain patterns are matched. Such a JAPE rule can be seen in Figure 9. The rule searches for instances of the child–parent relationship by looking for specific linguistic construction patterns. The presented rule,

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

```
Phase: Kinship_XSonOfY
Input: Lookup Token SpaceToken TempPerson Title
                                         TitleComplex
Options: control = appelt

Macro: PERSON_COLLECTION
(
  ({TempPerson}
     ({Token.kind == punctuation, Token.string == ","} |
       (SPACE {Token.string == "si"}))?
     SPACE )+
  {TempPerson}
)

Macro: PERSON_COMPLEX
(
  {TempPerson}
  ({Token.kind == punctuation, Token.string == ","} |
     {Token.string == "si"})?
  SPACE
  ({TitleComplex} | {Title})
)

Rule: CandidateKinship_XSonOfY
(
   ((PERSON_COLLECTION) |
     PERSON_COMPLEX |
     {TempPerson})
   ({Token.kind == punctuation, Token.string == ","})?
  SPACE
  {Lookup.majorType == kinship_relations}
  SPACE
  ({Token.string == "lui"}
     SPACE)?
  {TempPerson}
):kinship -->
      :kinship.Kinship_XSonOfY =
```

Figure 9: Example of a JAPE rule

Carol Robert, regele Ungariei, doneaza lui Mihail si Nicolae, fiii lui Albert de Juk mosia Palostelek si padurea Imbuz (Omboz), din comitatul Dabaca, pentru serviciile de arme credincioase aduse alaturi de magistrul Stefan, impotriva lui Moise, razvratit impotriva coroanei.

a)


```
- <paragraph>
  - <EventDonationComplex>
    - <PersonComplex>
        <Person>Carol Robert</Person>
      - <TitleComplex>
          <Title>regele</Title>
          <Location>Ungariei</Location>
        </TitleComplex>
      </PersonComplex>
      <EventDonation>doneaza</EventDonation>
    - <Kinship_XSonOfY>
      - <PersonCollection>
          <Person>Mihail</Person>
          <Person>Nicolae</Person>
        </PersonCollection>
        <Person>Albert de Juk</Person>
      </Kinship_XSonOfY>
    </EventDonationComplex>
  - <LocationTypeCollection>
      <LocationType_Estate>Palostelek</LocationType_Estate>
      <LocationType_Forest>Imbuz</LocationType_Forest>
    </LocationTypeCollection>
    <LocationType_County>Dabaca</LocationType_County>
    <Highlight>serviciile de arme credincioase</Highlight>
  - <PersonComplex>
      <Title>magistrul</Title>
      <Person>Stefan</Person>
    </PersonComplex>
  - <Antagonist>
      <Person>Moise</Person>
    </Antagonist>
    <EventRebellion>razvratit</EventRebellion>
  </paragraph>
```


b)

Figure 10: Information extraction results.

CandidateKinship_XSonOfY, finds phrasal patterns of the form “X son of Y”, in order to, annotate the lexical elements X and Y as Person, Person Collection (several persons connected by commas and conjunctions) or Complex Person (a lexical construct consisting of a name and a title).

For the identification of proper names in the raw documents, we used an existing Romanian gazetteer (Tablan et al., 2004), that provides lists of Romanian words. We have enriched the gazetteer with additional lists that contain information specific to the addressed historical periods, such as events, kinship relations, titles, estates, etc.

Within the process of annotating the lexical elements, the raw document is passed along with the gazetteer lists through the pipeline of JAPE grammars for extracting and structuring the relevant information. Inside this process, the JAPE grammars are used with the ANNIE (Tablan et al., 2004) information extraction system integrated in our system using the GATE API (Tablan et al., 2004).

For the raw document shown in Figure 7, the identified lexical elements are presented in Figure 10a. The result of the annotation of the lexical elements is the XML file shown in Figure 10b. It contains a hierarchic structure of the identified lexical elements that are further semantically annotated by using a set of mapping rules (see Figure 11). Each mapping rule associates to a specific pattern defined in the XML file *(i)* a set of ontology concepts that semantically annotate the lexical elements and *(ii)* a set of operations that need to be performed on the ontology in order to store the identified information (instance population, definition of properties and relations).

In the mapping rule shown in Figure 11, for the lexical tag *Kinship_XSonOfY*, the child elements X and Y are semantically annotated with the ontology concept *Person*. The mapping rule also specifies the actions of *(i)* adding X and Y as instances of *Person* into the ontology and *(ii)* defining the *hasFather* relation between the two instances.

For the currently processed raw document, the

```
<rule id="10">
   <sem_tag>Kinship_XSonOfY</sem_tag>
   <context>
      <child_tag>Person</child_tag>
      <child_tag>Person</child_tag>
   </context>
   <actions>
<action>
         <atype>addInstance</atype>
         <aclass>Person</aclass>
         <aobject child="yes" index="0">Person</aobject>
      </action>
      <action>
         <atype>addInstance</atype>
         <aclass>Person</aclass>
         <aobject child="yes" index="1">Person</aobject>
      </action>
      <action>
         <atype>hasFather</atype>
         <aclass>Person</aclass>
         <aobject index="1">Person</aobject>
      </action>
   </actions>
</rule>
```

Figure 11: Example of a mapping rule.

knowledge extraction process also generates an RDF file (see Figure 12) that contains RDF statements capturing the semantic annotations of the document.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:features="http://edu.utcluj.arhinet/features#"
  xmlns:ontClass="http://edu.utcluj.arhinet/onto#" >
 <rdf:Description rdf:about="http://edu.utcluj.arhinet/features#DocumentFeatures">
  <features:docOriginalPath>C:\Arhinet\corpus\testFile1.txt</features:docOriginalPath>
  <features:docAnnotatedPath>C:\Arhinet\rdf\testFile1.rdf</features:docAnnotatedPath>
 </rdf:Description>
 <rdf:Description rdf:about="http://edu.utcluj.arhinet/onto#DocumentAnnotation">
  <ontClass:Event>razvratit</ontClass:Event>
  <ontClass:Title>magistrul</ontClass:Title>
  <ontClass:Event>doneaza</ontClass:Event>
  <ontClass:Title>regele</ontClass:Title>
  <ontClass:Person>Carol Robert</ontClass:Person>
  <ontClass:Location>Ungariei</ontClass:Location>
  <ontClass:Person>Stefan</ontClass:Person>
  <ontClass:Person>Moise</ontClass:Person>
 </rdf:Description>
</rdf:RDF>
```

Figure 12: Example of RDF file.

After processing several documents within the knowledge acquisition workflow, the domain ontology is populated with new instances and properties (see Figures 8 and 13).

Figure 14 illustrates an example of a semantic query that aims to search all documents that provide information about the territorial division "Palostelek". Most of the archival documents related to Transylvania may contain names and terms written in different forms because of linguistic and phonetic influences (Romanian, Hungarian, German or Slavic). Consequently, the location name "Palostelek" may also appear as "Paulusteleky". Similarly, the location name "Juc" may appear as "Szuc", "Suzuluk", "Zuk" or "Suk".

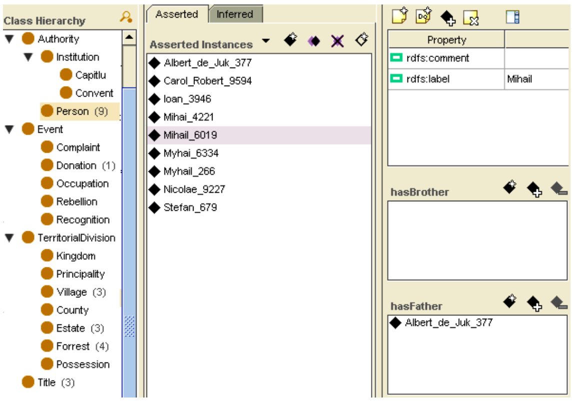


Figure 13: Ontology population results.

```
TerritorialDivision(?t0) ^ rdfs:label(?t0, ?t0_name)
   ^ swrlb:contains(?t0_name, "Palostelek")
   ^ sameAs(?t0,?t0s)  ^  rdfs:label(?t0s, ?t0s_name)
   -> sqwrl:selectDistinct(?t0s_name, TerritorialDivision")
```

Figure 14: Query example.

When performing queries about a term, like "Palostelek", it is essential to also identify all documents about "Paulusteleky" and other possible formats. This problem is solved within the synonym search functionality implemented as part of the knowledge processing and retrieval workflow (see Section 6).

The whole set of relevant document contents obtained after executing the query illustrated in Figure 14 is presented in Figure 15.

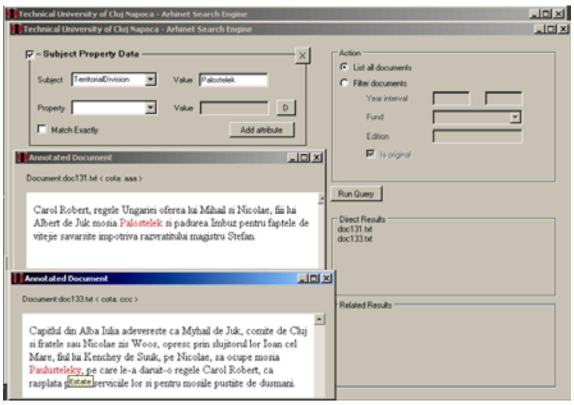


Figure 15: Query results.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

8 CONCLUSIONS AND FUTURE WORK

The present paper proposes a generic model of the archival domain and offers a technical solution for generating, semantically enhancing and processing archival eContent. The solution follows two main workflows: *knowledge acquisition* and *knowledge processing and retrieval*. Within the knowledge acquisition workflow we extend the OntoPop methodology by adding two processing steps that regard synonym and homonym population. Reasoning techniques applied in knowledge processing and retrieval enable ontology-guided intelligent queries aiming at the identification of relevant documents and knowledge.

For future work, we intend to extend our system to enable ontologically-guided natural language query processing and multilingual transparency. Moreover, the domain ontology will be automatically extended with new knowledge extracted from documents.

ACKNOWLEDGEMENTS

This work was supported by the ArhiNet project within the framework of the “Research of Excellence” program initiated by the Romanian Ministry of Education and Research.

REFERENCES

Amardeilh, F., 2007. Web Sémantique et Informatique Linguistique: propositions méthodologiques et réalisation d’une plateforme logicielle. These de Doctorat, Universite Paris X-Nanterrere.

Amardeilh, F., 2006. OntoPop or how to annotate documents and populate ontologies from texts. in *Proceedings of the ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation*, Budva, Montenegro, June 12, 2006. *CEUR Workshop Proceedings,*
ISSN 1613-0073.

Buitelaar, P., Cimiano, P., Racioppa S., Siegel, M., 2006. Ontology-based Information Extraction with SOBA. In *Proceedings of the International Conference on Language Resources and Evaluation,* pp. 2321-2324.

Laclavik M., Ciglan M, Seleng M, Krajei S., 2007. Ontea: Semi-automatic Pattern based Text Annotation empowered with Information Retrieval Methods. In *Tools for acquisition, organisation and presenting of information and knowledge: proceedings in Informatics and Information Technologies,* Kosice: Vydavatelstvo STU, Bratislava. ISBN 978-80-227-2716-7, part 2, pp. 119-129.

Schäfer, U., 2007. Integrating Deep and Shallow Natural Language Processing Components – Representations and Hybrid Architectures. Saarbrücken Dissertations in Computational Linguistics and Language Te, DFKI GmbH and Computational Linguistics Department, Saarland University, Saarbrücken, Germany.

Tablan V., Maynard D., Bontcheva K., Cunningham H., 2004. Gate – An Application Developer’s Guide. Available online: http://gate.ac.uk/.

Sandia National Laboratories, 2008. Jess the Rule Engine for the Java Platform, Version 7.1. Available online: http://www.jessrules.com/jess/docs/Jess71.pdf.

Horrocks I., et al., 2004. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. Available online: http://www.w3.org/Submission/SWRL/.

SQWRL: Semantic Query-Enhanced Web Rule Language. Available online: http://protege.cim3.net/cgi-bin/wiki.pl?SQWRL, Date accessed 01/06/2008.

Horridge, M., et al., 2007. A Practical Guide to Building OWL Ontologies Using Protégé 4 and CO-ODE Tools. Available online: http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial-p4.0.pdf.

CCNA, Cluj County National Archives, 2008,
Online: http://www.clujnapoca.ro/arhivelenationale/.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS