Project title: Integrated System for developing semantically-enhanced archival content ARHINET

Main objective: The development of an experimental model of an integrated system for creating, managing and processing digital content extracted from archival sources using ontologies and semantic annotations.

Introduction

The topic of this interdisciplinary project can be integrated in the broader research field of **Interactive digital content management systems (e-content).**
The objective of the project is the development of an experimental model of an integrated system for creating and managing archival content using ontologies and semantic annotations. Based on the created content that is semantically annotated, the systems allows for relevant information retrieval.

Conceptual architecture of the integrated system

To achieve the above mentioned objectives we developed an integrated three-layered architecture: (i) data acquisition and annotation, (ii) knowledge processing and (iii) querying and information retrieval.

We present in Fig.1 the conceptual architecture of the integrated system..

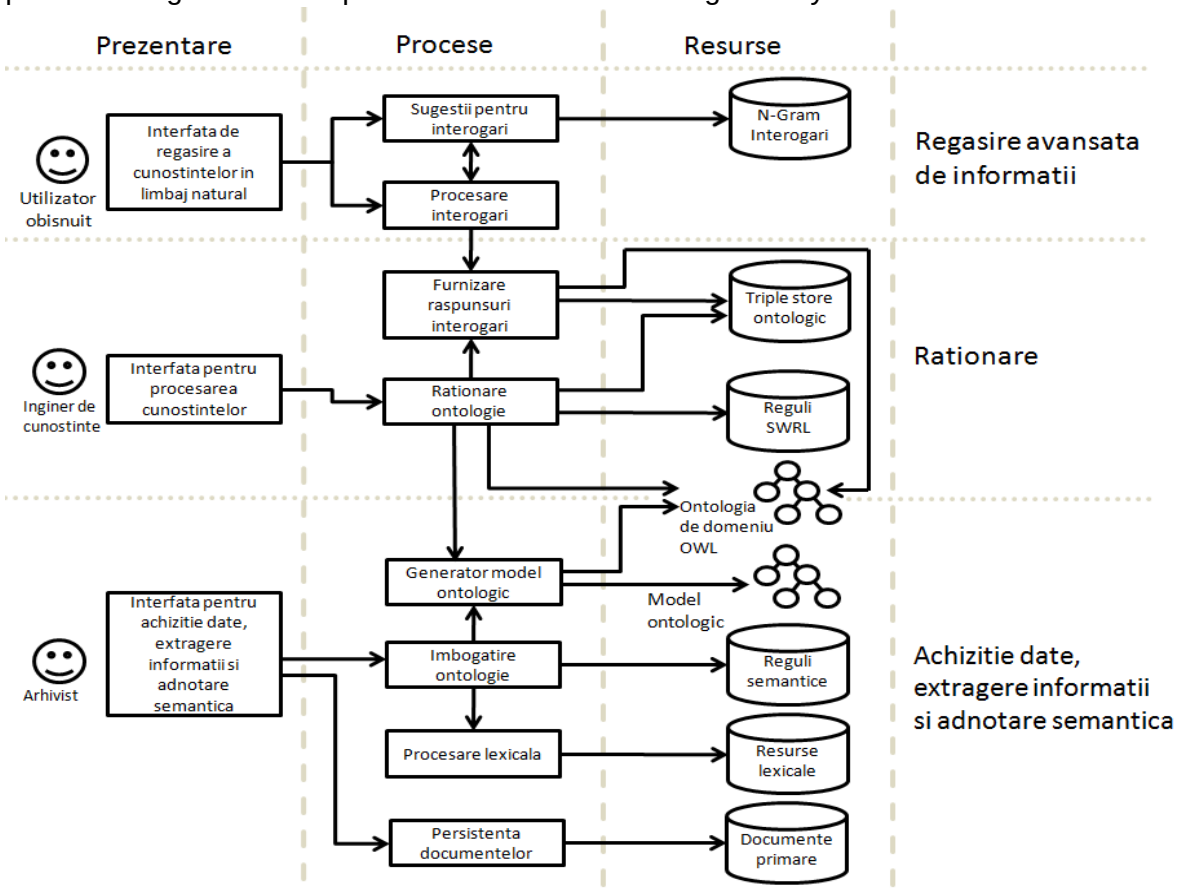
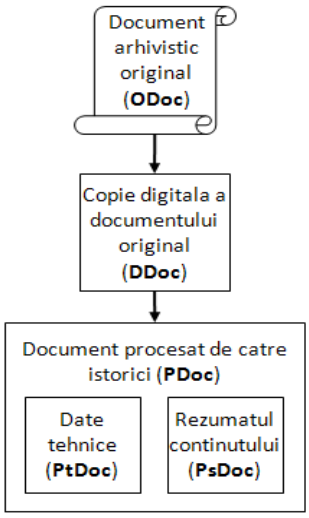


Figure 1 Conceptual architecture of the integrated system

Data acquisition and annotation

For the development of the **data acquisition and annotation** layer, we identified the data sources, the algorithms for automatically importing the data and we defined the processing flow for data input and annotation. Considering how the archival data is structured and stored in the National Archive storage we designed the data model and the database structure. Based on the designed models we developed an experimental model for data acquisition and annotation.

The primary data is obtained by pre-processing the original archival documents (named ODoc) that address the medieval history of Transylvania. The original documents are heterogenous in: the language the document was written (romanian, latin, hungarian or german), the institution that issued the document, the way the document was written (printed or hand written), existence of letter embelishments.These features hinder the automatic document processing so we decided to usa as primary data the existing summaries obtained after the original documents were digitized (DDoc) and processed. After processing a new document is created (Pdoc) that contains both technical data (PtDoc) and the summary of the original content. The technical data is related to the issue date of the document, the physical features (document support, physical state of the support, etc). Fig. 2 presents the preprocessing steps that lead to the Pdoc with the two components: PtDoc and PsDoc.



To achieve the information retrieval and semantic annotation objectives, we developed lexical and semantic annotation processes. These processes are implemented using GATE functionalities provided via an API that allows accessing and adapting the functionalities according to the project needs. Thus, the lexical annotation is based on on the components of the GATE pipeline while the semantic annotation is based on JAPE rules. The annotation processes are followed by the domain ontology enrichment For this purpose we used the Text2Onto framework that includes several algorithms for extracting modeling primitives from the annotations performed in the previous steps. These primitives are translated in OWL DL, the ontology representation language and stored in the ArhiNet database. We present in Fig.3. the architecture of the information extraction and semantic annotation system.

Figure 2 Original documents preprocessing steps

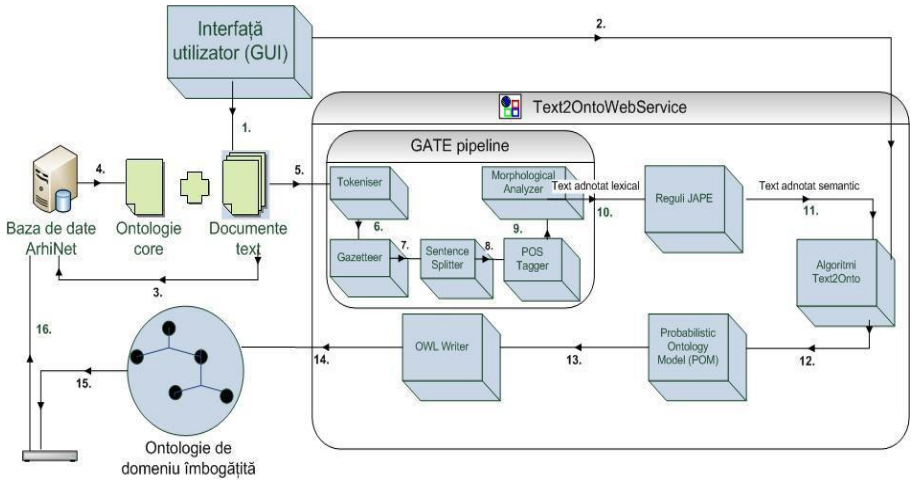


Figure 3 Architecture of the information extraction and semantic annotation system

Knowledge processing

The **knowledge processing** layer involved the development of a formal knowledge representation model. In our model the terminological component TBox is represented as an ontology and the asertional component ABox is represented in a relational data model as triplets (Subject, Property, Object). An associated set of SWRL rules are used to infer and generate new pieces of knowledge.

The Ontology management activities include: (*i*) ontology consistency check, (*ii*) ontology classification, (*iii*) ontology realization, (*iv*) rule-based inference and (*v*) SPARQL queris processing.

The overall architecture of the knowledge processing system is presented in fig.4.

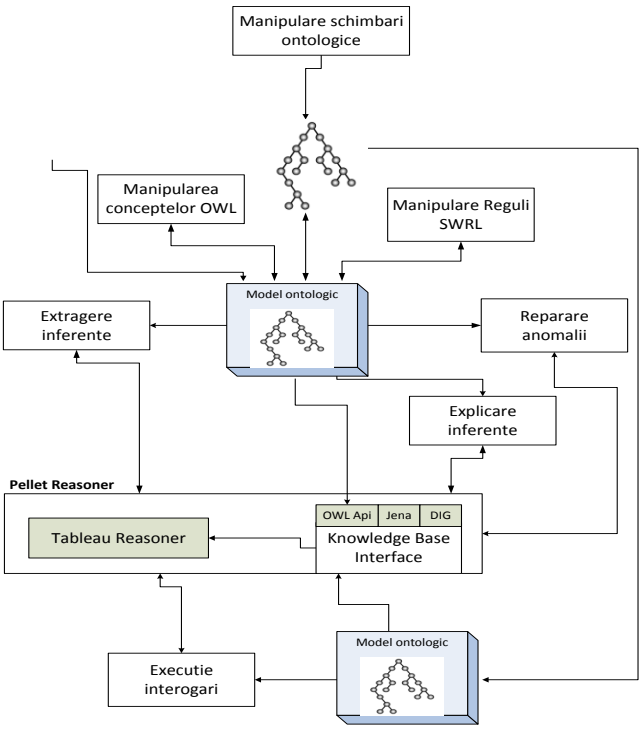


Figure 4 Knowledge processing

Querying and Information retrieval

Using the **querying and information retrieval** layer the user is able to identify information and historical documents by entering ontology-guided natural language queries. The system provides a set of suggestions that are relevant in the query context. The suggestions are based on a query grammar that describes the structures of possible questions in the romanian language and on the domain ontology.

The system is structured in three modules: (1) the Suggestions Module that deals with extracting suggestions from the domain ontology, (2) the Query Processing Module that translates the queries provided by the user using natural language in SPARQL queries that are processed by the knowledge processing system, (3) the Results Processing Module that handles the way the results are displayed and identifies the corresponding documents to each result. The architecture of this system is presented in fig.5.

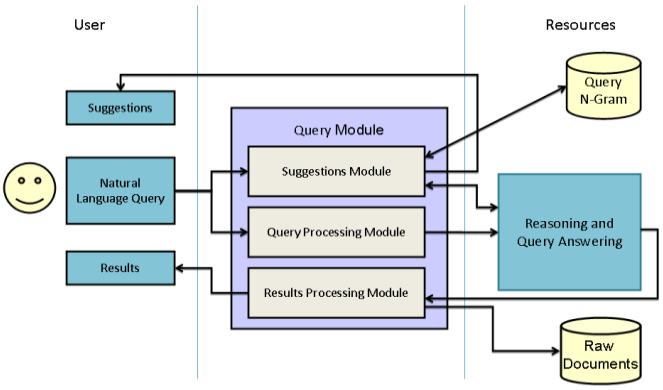


Figure 5 Information retrieval