

Titlu proiect: Sistem integrat de dezvoltare de conținut arhivistic adnotat semantic - ARHINET

Obiectivul principal: Dezvoltarea unui model experimental de sistem integrat de generare, gestiune și procesare de conținut digital provenit din surse arhivistice pe baza adnotărilor semantice.

Introducere

Tema proiectului are un caracter interdisciplinar și se încadrează în direcția de cercetare **1.2.6**

Sisteme informatice pentru gestiunea și dezvoltarea de conținut digital interactiv (e-content).

Obiectivul proiectului este dezvoltarea unui model experimental al unui sistem integrat de generare, gestiune și procesare de conținut digital provenit din surse arhivistice folosind ontologii de domeniu și adnotări semantice. Pe baza conținutului generat și adnotat se pot regăsi informații relevante semantic.

Arhitectura conceptuală a sistemului integrat

Pentru realizarea obiectivelor, s-a dezvoltat o arhitectură integrată, pe trei nivele: (i) achiziția, organizarea și adnotarea datelor, (ii) procesarea cunoștințelor și (iii) interogarea și regăsirea informației.

În fig.1. este prezentată arhitectura conceptuală a sistemului integrat.

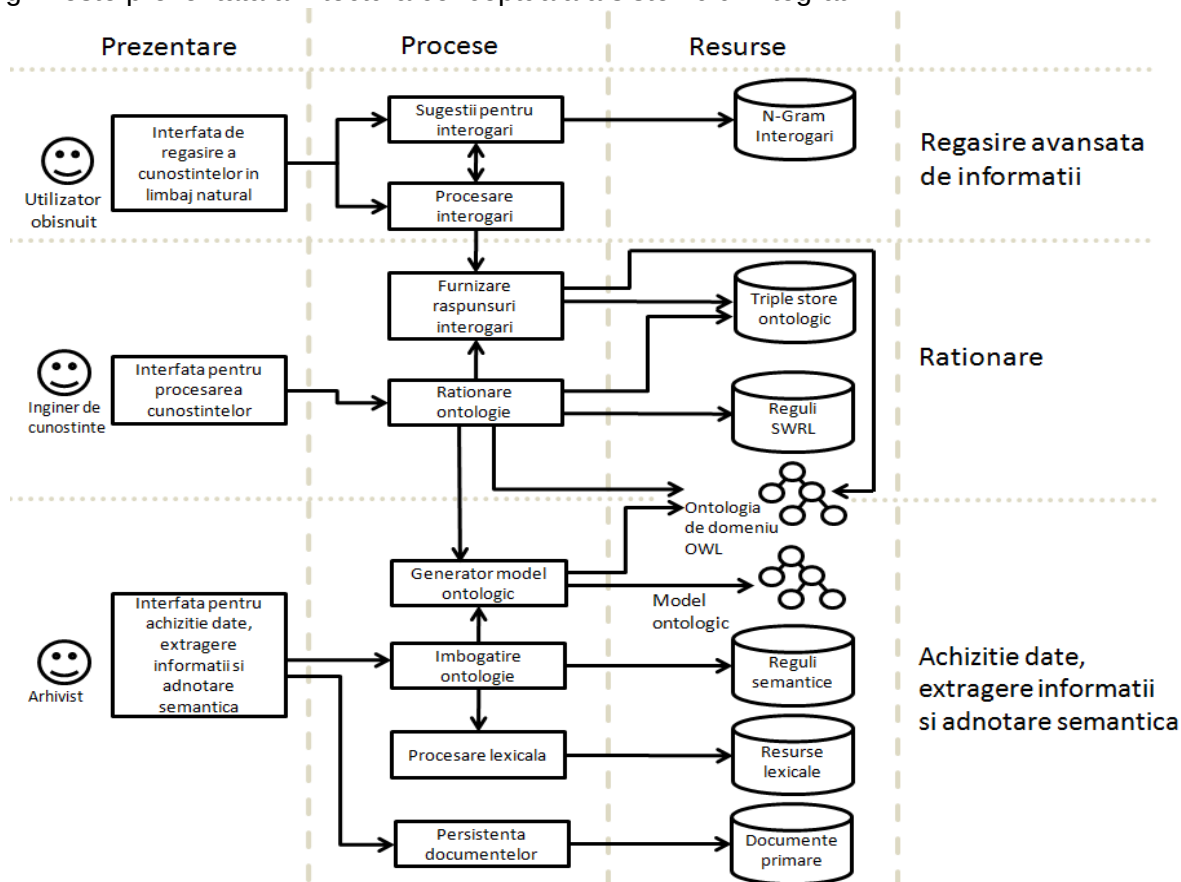


Figure 1 Arhitectura conceptuală a sistemului integrat

Achizitia si adnotarea datelor

Pentru realizarea **achizitiei si adnotarii datelor** din surse arhivistice s-au identificat sursele de date, modalitatea de importare automată a datelor și s-a definit fluxul de introducere al datelor și de adnotare semantică. Pornind de la modalitatea de organizare a datelor primare în cadrul arhivelor s-a definit modelul datelor si structura bazei de date primare. Pe baza modelelor proiectate s-a dezvoltat un model experimental de achizitie si adnotare a datelor.

Datele primare sunt obtinute prin preprocesarea unui set de documente arhivistice originale (denumite documente ODoc) care vizeaza istoria medievala a Transilvaniei. Documentele originale sunt heterogene prin: limba in care a fost scris documentul (romana, latina, maghiara sau germana), institutia care a emis documentul, forma in care a fost scris documentul (tiparit sau scris de mana), ornamentele aduse literelor. Aceste caracteristici ingreuneaza procesarea automata a documentelor si de aceea am hotarat sa folosim ca si date primare pentru sistem rezumatele documentelor obtinute dupa digitizare (DDoc) si procesare. In urma procesarii se creeaza un nou document (PDoc) care contine datele tehnice (PtDoc) si rezumatul documentului original (PsDoc). Datele tehnice se refera la data eliberarii documentului, caracteristicile fizice (materialul pe care a fost scris, starea fizica a documentului, etc.) ale documentului.

Figura 2 prezinta pasii de preprocesare care creeaza documentul PDoc cu componentele PtDoc si PsDoc.

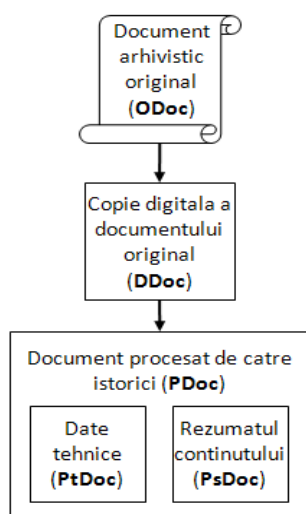


Figure 2 Pasii de procesare a documentelor istorice

Pentru atingerea obiectivelor modelului experimental al sistemului de extragere a informatiilor si adnotare semantica, sunt implicate procese de adnotare lexicală si semantică a documentelor. Aceste procese sunt realizate folosind functionalitățile oferite de GATE si puse la dispozitie printr-un API, care permite utilizarea si adaptarea componentelor sale, conform necesităților proiectului. Astfel, adnotarea lexicală a documentelor se bazează pe componentele din fluxul GATE, iar adnotarea semantica se bazeaza pe reguli JAPE. Procesele de adnotare sunt urmate de îmbogățirea ontologiei de domeniu. In acest scop am utilizat framework-ul Text2Onto, care prin diversi algoritmi permite extragerea de primitive de modelare din adnotările făcute la pasul anterior. Aceste primitive de modelare sunt ulterior translate în limbajul de reprezentare a ontologiilor OWL DL, si stocate în baza de date ArhiNet, îmbogățind astfel baza de cunostinte, i.e. ontologia de domeniu a sistemului. In fig. 3 este prezentata arhitectura sistemului de extragere a informatiilor si adnotare

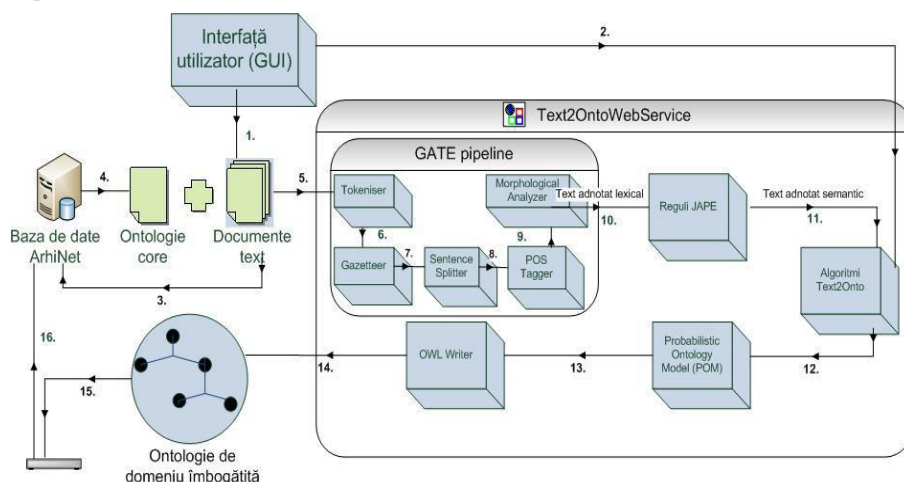


Figure 3 Arhitectura sistemului de extragere informatii si adnotare semantica

Procesarea cunostintelor

Pentru procesarea cunostintelor s-a dezvoltat un model formal de reprezentare a cunostintelor. In acest model componenta terminologica TBox este reprezentata sub forma unei ontologii, iar componenta asertionala ABox este reprezentata intr-un model de date relational sub forma unor triplete (subiect, proprietate, obiect). Un set de reguli SWRL asociate sunt utilizate pentru inferarea si generarea de noi cunostinte. Managementul ontologiei include urmatoarele operatii: (i) verificarea consistentei ontologiei, (ii) clasificarea ontologiei, (iii) realizarea ontologiei, (iv) inferenta bazata pe reguli si (v) procesarea interogarilor in SPARQL.

Arhitectura sistemelor de procesare a cunostintelor este prezentata in fig.4.

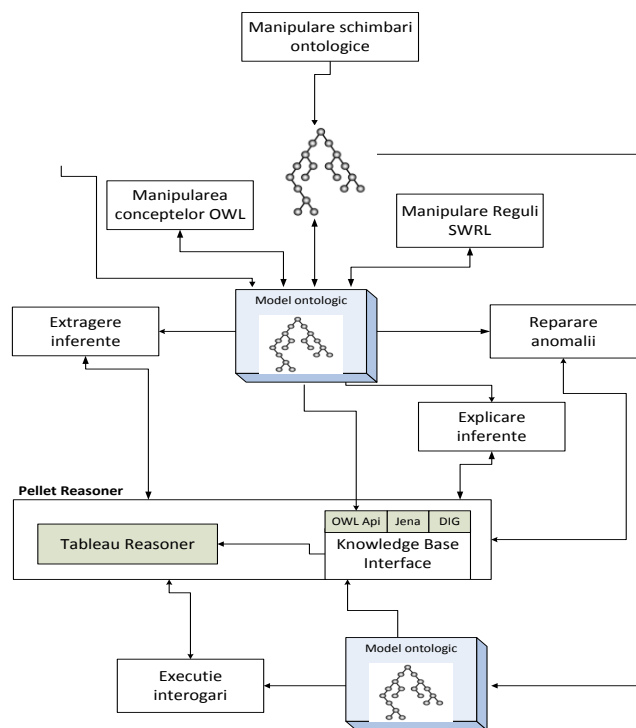


Figure 4 Procesarea cunostintelor

Interogarea si regasirea informatiilor

Sistemul de **interogare si regasire** avansata a informatiilor permite utilizatorilor sa identifice informatii si documente istorice prin intermediul unor interogari in limbaj natural ghidate ontologic prin care utilizatorului i se furnizeaza un set de sugestii relevante contextului de interogare. Aceste sugestii sunt formulate pe baza unei gramatici de interogari care descrie structurile de intrebari posibile pentru limba romana si elemente din ontologia de domeniu. Sistemul este impartit in trei module: (1) modulul de sugestii (Suggestions Module) care se ocupa de colectarea sugestiilor din ontologia de domeniu, (2) modulul de procesare a interogarilor (Query Processing) care translateaza interogarile utilizatorilor scrise in limbaj natural in interogari SPARQL ce vor fi procesate de sistemul de rationare, (3) modulul de procesare a rezultatelor (Results Processing) care se ocupa de afisarea rezultatelor si identificarea documentelor corespunzatoare fiecarui rezultat. Arhitectura acestui sistem este prezentata in fig.5.

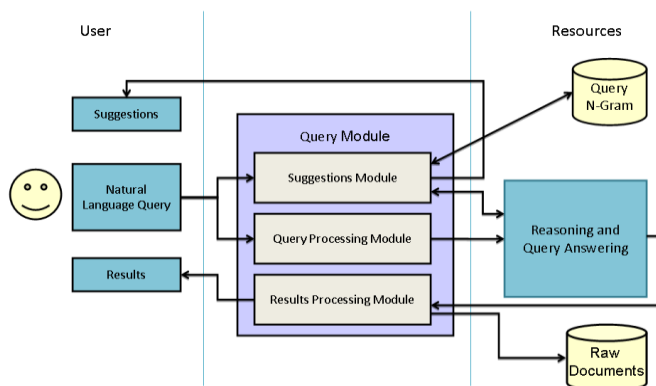


Figure 5 Regasirea informatiei