Barometr Regionalny
Nr 3(21) 2010

ControlSem – system wspierania decyzji oparty na technologiach sieci semantycznych

Jan Andreasik
Wyższa Szkoła Zarządzania i Administracji w Zamościu
Instytut Informatyki Biomedycznej Wyższej Szkoły Informatyki i Zarządzania w Rzeszowie

Andrzej Ciebiera
Instytut Informatyki Biomedycznej Wyższej Szkoły Informatyki i Zarządzania w Rzeszowie

Sławomir Umpirowicz
Instytut Informatyki Biomedycznej Wyższej Szkoły Informatyki i Zarządzania w Rzeszowie
Partners in Progress

Streszczenie: *Duża ilość danych przechowywana w bazach relacyjnych nastręcza trudności w ich przeglądaniu, wyszukiwaniu i przygotowywaniu raportów. Zapytania do baz danych wykonywane na powielonych, niespójnych i bezkontekstowych danych powodują generowanie błędnych wyników. Systemy oparte na rozwiązaniach sieci semantycznych (Semantic Web) pozwalają na budowę systemów opartych na wiedzy (Knowledge Based Systems). Niniejszy artykuł przedstawia architekturę systemu ControlSem opracowanego dla podkarpackiego oddziału Narodowego Funduszu Zdrowia – systemu dedykowanego dla kontroli procedur medycznych opartego na technologiach SW. Doświadczenia z projektowania systemu i modelowania ontologii dziedzinowej oraz procesu przetwarzania danych mogą być wykorzystane w implementacjach rozwiązań kontekstowej integracji i analizy danych.*

Wprowadzenie

W ostatnich latach nastąpił intensywny rozwój technologii sieci semantycznych SW (*Semantic Web*) w kierunku semantycznej analizy danych z dużych repozytoriów [*Oracle Database*… 2009]. W wielu instytucjach i przedsiębiorstwach funkcjonują i są nadal rozszerzane duże, rozproszone bazy danych oparte na technologiach relacyjnych baz danych firm ORACLE, MICROSOFT i innych. Przegląd danych nastręcza dużych trudności. Związane są one z niedostosowaniem wielu technologii przeglądu rozproszonych baz danych i raportowania do wymagań użytkownika. Użytkownikami baz danych nie są wyspecjalizowani informatycy znający biegle języki zapytań (SQL), ale pracownicy administracyjni, specjaliści działów marketingu i finansów oraz kontrolerzy w różnorodnych systemach controllingu. Specjaliści ci chcą formułować zapytania w naturalnym języku danej dziedziny przedmiotowej. Chcą również uzyskiwać nie tylko raporty z danymi czy też informacjami tworzonymi na podstawie relacji pomiędzy danymi. Chcą pozyskiwać wiedzę dotyczącą określonej problematyki, tworząc tym samym system oparty na wiedzy KBS (*Knowledge Based System*). Takie możliwości daje im technologia SW (*Semantic Web*) [*Semantic Technologies*… 2009].

Mechanizm wnioskowania sieci semantycznych jest oparty na dopasowaniu reprezentacji zapytania do ontologii sieci semantycznej, która stanowi system informacyjny i w której wszelkie powiązania pomiędzy obiektami sieci są przewidziane i w sposób jawny zdefiniowane. Organizacja W3C (World Wide Web Consortium) opracowała szereg standardów, w szczególności bazujących na składni dokumentów XML, na których oparte są technologie sieci semantycznych:

1. RDF (*Resource Description Framework*) – standard opisu danych zawierający informacje w postaci grafu skierowanego, oparty na trójce metadanych (podmiot, właściwość, wartość właściwości), pozwalający na maszynowe przetwarzanie zasobów opisanych w sposób abstrakcyjny;

2. RDFS (*Resource Description Framework Scheme*) – język reprezentacji wiedzy pozwalający na strukturalne uporządkowanie wiedzy opisanej przez trójki RDF;
3. OWL (*Web Ontology Language*) – język pozwalający na budowę ontologii opisujących w sposób formalny zbiory obiektów (podmiotów) i zależności pomiędzy nimi.

W niniejszym artykule przedstawiona zostanie architektura systemu ControlSem opartego na technologiach sieci semantycznych, opracowanego dla podkarpackiego oddziału Narodowego Funduszu Zdrowia. Zdefiniowano listę problemów obejmujących proces przeszukiwania rozproszonych bazy danych NFZ. Problematyka dotyczy kontroli procedur medycznych. Kontrolerzy NFZ chcą uzyskać raporty pod określone, skomplikowane zapytania.

Tego typu systemy dopiero są tworzone. Można wymienić tu kilka systemów opartych na repozytoriach szpitali, przy czym tam problematyka obejmuje procesy monitorowania badań pacjentów. Taki system HIWO (*Hospital Intelligent Ward Ontology*) przedstawili P. Kataria, R. Juric, S. Paurobally, K. Madani [Kataria i in. 2008]. Ontologię HIWO przedstawiono w narzędziu TopBraid[1], natomiast konwersji danych z relacyjnej bazy Oracle Express Edition 10g do formatu RDF dokonano za pomocą narzędzia D2RQ, które wykorzystuje biblioteki Jena API [*The D2RQ*…]. Innym podejściem jest opracowanie P. LePendu, D. Dou, G.A. Frishkoff, J. Rong [LePendu i in. 2008], którzy utworzyli bazę ontologiczną do analizy elektroencefalografów. Dane gromadzono w bazie MySQL RDBMS, reguły zapisano w języku SWRL[2], a zapytania skonstruowano w językach: SPARQL[3], OWL-QL[4], SQL[5]. Zespół pracowników H. Chen [Chen i in. 2006] opracował narzędzie Dartgrid do przetworzenia danych z relacyjnej bazy, posługując się językiem zapytań SPARQL. Opracowano aplikację dotyczącą analizy procedur medycyny chińskiej, bazując na bazie China Academy of Traditional Chainise Medicine CTCM.

Innym podejściem jest tworzenie systemów zaprojektowanych zgodnie z zasadami technologii SW. Takie systemy oparte są na medycznych ontologiach, takich jak UMLS[6], SNOMED[7], GALEN[8]. System zapisu procedur medycznych w języku OWL przedstawili A.L. Rector, R. Qamar, T. Marley [Rector i in. 2009]. System zarządzania informacją medyczną w szpitalu oparty na językach OWL-S, OWL, SWRL przedstawili M.A. Casteleiro, J.J. Des Diz [Casteleiro i in. 2008]. Opracowali oni interfejs pomiędzy tezaurusem UMLS a edytorem języka OWL, Protege-OWL[9].

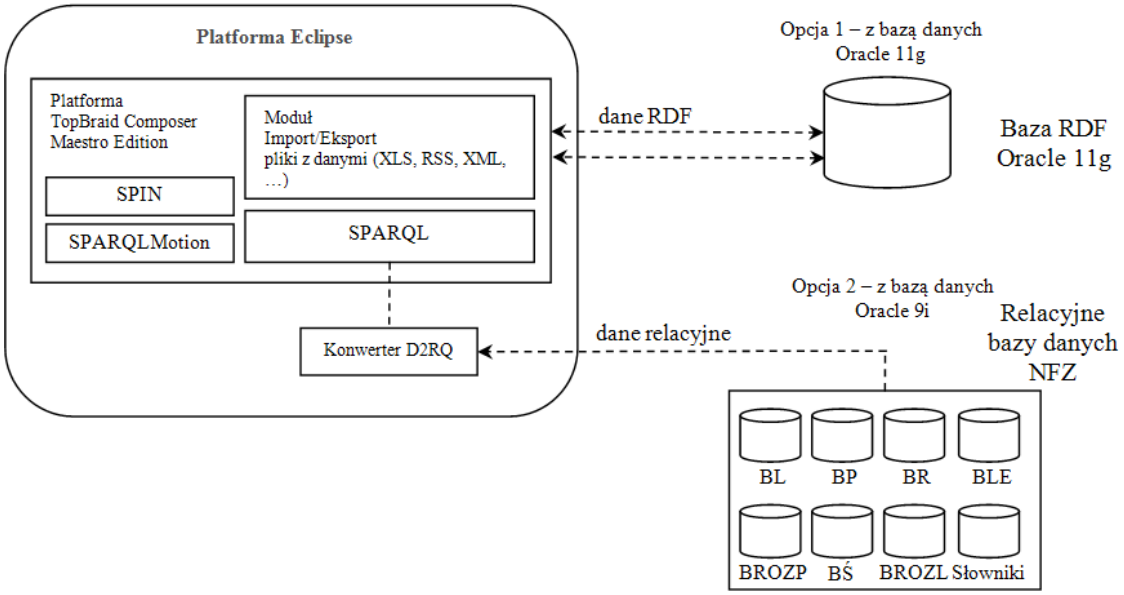
1. Architektura systemu

Zastosowanie SW wymaga podłączenia baz danych pozwalających na przechowywanie modelu RDF albo zaimportowanie danych do struktur RDF. Dane źródłowe dla ControlSem są przechowywane przez NFZ w relacyjnych bazach danych firmy Oracle. Z tego względu do podłączenia źródeł danych wykorzystaliśmy konwerter D2RQ. Jako narzędzie modelowania SW zostało wykorzystane narzędzie TopBraid Composer (TBC) – rozbudowany edytor języków RDF, RDFS i OWL, a także eksplorator instancji i narzędzie do wykonywanie zapytań SPARQL (odpowiednika języka zapytań SQL dla relacyjnych baz danych) z graficznym interfejsem użytkownika. Dodatkową zaletą TBC jest wbudowany silnik regułowy SPIN (SPARQL Inferencing Notation), który znacznie ułatwia definiowanie dodatkowych warunków i reguł w konstruowanych zapytaniach systemu ControlSem.

W przyszłości przy zastosowaniu bazy Oracle 11g możliwa będzie rezygnacja z narzędzi D2RQ. Zbiór relacyjnych baz danych NFZ składa się ze zbioru kilkunastu baz. Do eksperymentu wybrano trzy bazy: baza leków, baza świadczeń i baza recept. Każda z baz cechuje się dużą liczbą rekordów (ok. 50 mln rekordów).

W dotychczas realizowanej przez podkarpacki oddział NFZ analizie danych opartej na zapytaniach SQL zidentyfikowano niespójności w danych źródłowych. Jednym z celów projektu było również potwierdzenie przydatności przyjętego modelu ControlSem w zakresie integracji baz.

[1] http://www.topquadrant.com.
[2] http://www.w3.org/Submision/2004/SUBM-SWRL-20040521/.
[3] http://www.w3.org/TR/rdf-sparql-query/.
[4] http://www.w3.org/TR/owl-features/.
[5] http://pl.wikipedia.org/wiki/SQL.
[6] http://www.nlm.nih.gov/research/umls/.
[7] http://www.ihtsdo.org/snomed-ct/.
[8] http://www.opengalen.org/.
[9] http://protege.stanford.edu/.



Rys. 1. Architektura systemu ControlSem

Dane źródłowe dla ControlSem zostały zidentyfikowane w postaci:

- Bazy lekarzy (BL),
- Bazy pacjentów (BP),
- Bazy recept (BR),
- Bazy leków (BLE),
- Bazy rozpoznań (BROZP),
- Bazy świadczeń (BŚ),
- Bazy rozliczeń (BROZL),
- Słownika międzynarodowych nazw leków (SLEK_SL_NAZWYM),
- Słownika ICD 9 (Międzynarodowa Klasyfikacja Procedur Medycznych),
- Słownika ICD 10 (Międzynarodowa Statystyczna Klasyfikacja Chorób i Problemów Zdrowotnych).

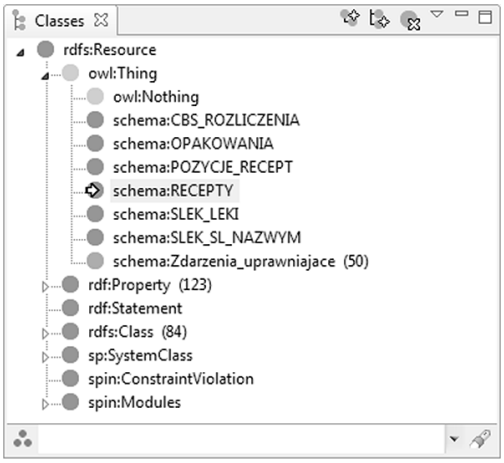
2. Ontologia ControlSem

Ze względu na fakt, iż ControlSem był projektowany do przetwarzania już istniejących źródeł danych, przyjęto organizację klas w zakresie struktur i właściwości w oparciu o struktury danych źródłowych zapisanych w bazie danych Oracle. Takie rozwiązanie skomplikowało implementację klas i ich właściwości w OWL, ale ułatwia pracę lekarzom specjalistom (znajomość dotychczasowych struktur) oraz zapewnia niezmienność organizacji istniejących źródeł danych niezbędną dla innych zastosowań.

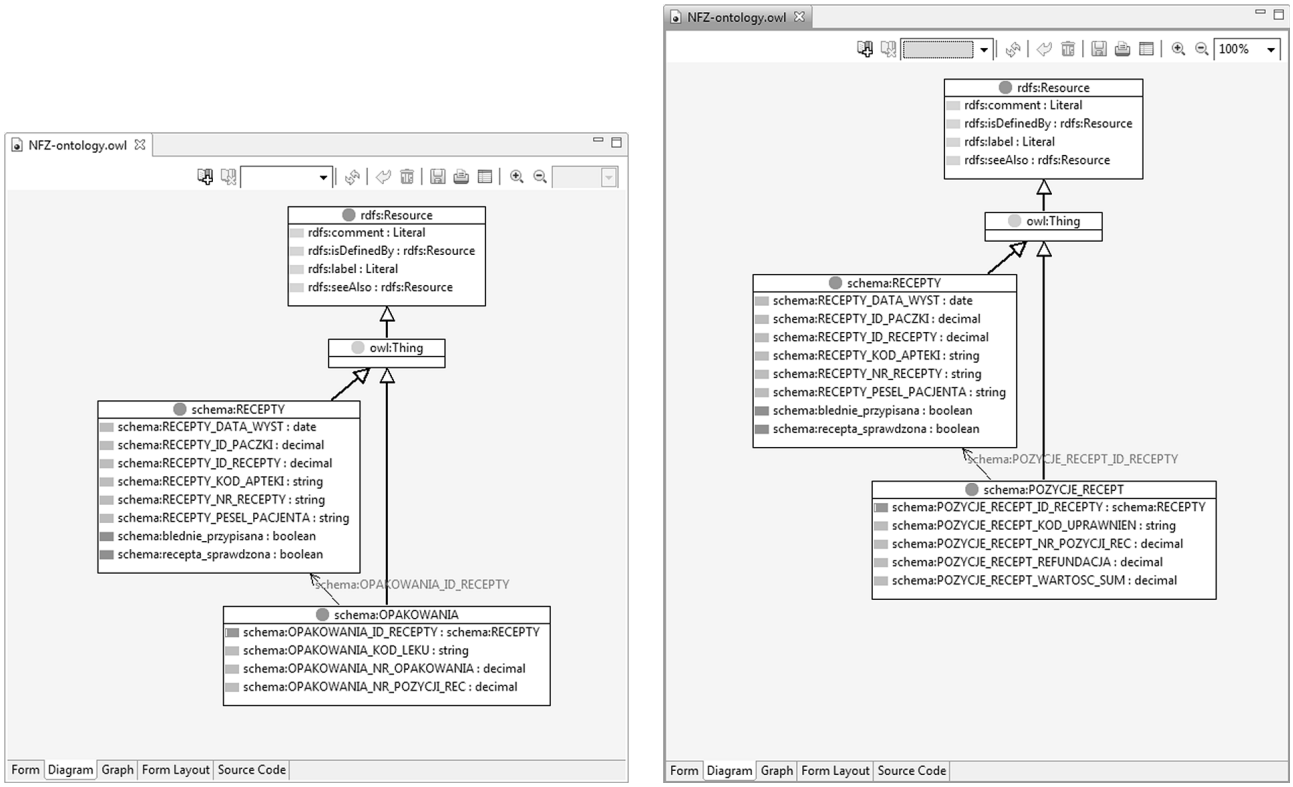
Dla potwierdzenia przydatności technologii SW w analizie procedur medycznych zostały zdefiniowane dwa rzeczywiste problemy – grupy procedur medycznych, w których stosowane są leki:
i – Tramadol i Pancreatinum,
ii – Clopidogrel.

Powyższe leki mogą być przepisywane na receptach w określonych procedurach medycznych (Proc) wyłącznie z odpowiednim kodem uprawnień (KodUpr) i są lekami o nadzorowanej dystrybucji. Różnica pomiędzy (i) oraz (ii) polega na tym, że przepisanie recepty na (i) nie jest uwarunkowane w czasie. Lek uznaje się za aplikowany prawidłowo w dowolnym czasie – przed, w czasie oraz po wystąpieniu zdarzenia uprawniającego (ZdUpr) do przepisania leku. Przepisanie recepty na (ii) to przypadek, w którym w zależności od rodzaju ZdUpr lek może być przepisywany wyłącznie w okresie właściwym dla rodzaju ZdUpr.

Definicja ontologii systemu ControlSem została przygotowana w edytorze TBC i jest przedstawiona na rys. 2.



Rys. 2. Diagram klas ControlSem w edytorze TBC

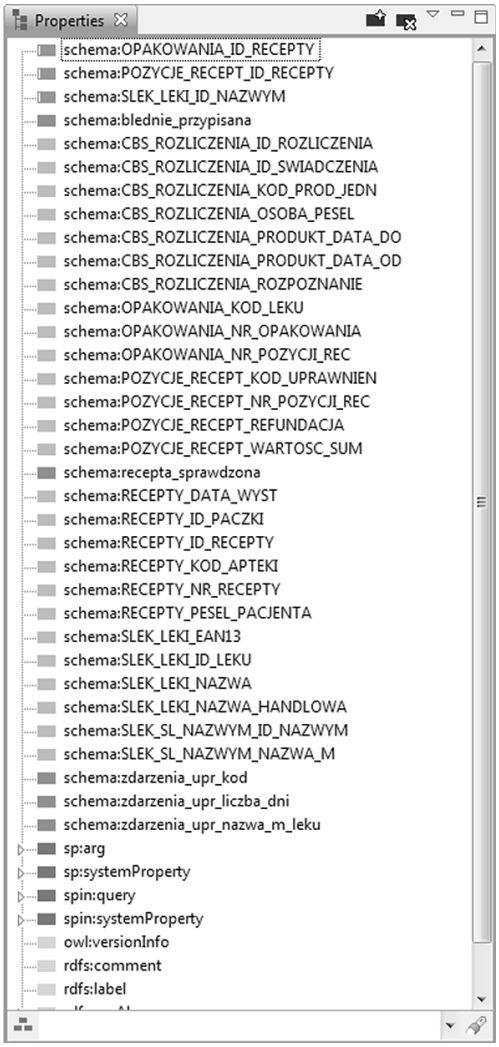


Rys. 3. Organizacja klas POZYCJE_RECEPT oraz OPAKOWANIA powiązanych z klasą RECEPTY

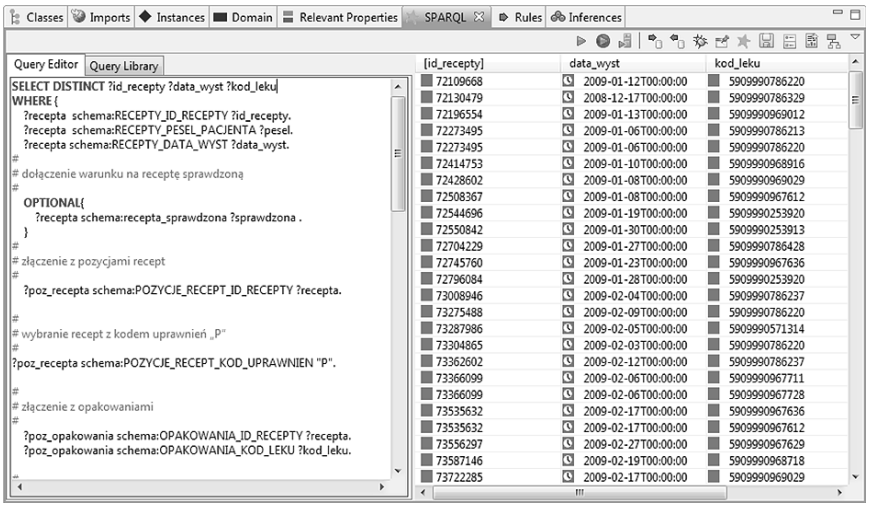
Przygotowana struktura odpowiada organizacji struktur aktualnie wykorzystywanych źródeł danych NFZ.

3. Analiza procedur medycznych z wykorzystaniem SPARQL i SPIN w ControlSem

Zdefiniowane problemy wnioskowania (i) oraz (ii) wymagają różnych ścieżek wyszukiwania. Dla przypadku (i) wyszukanie numerów recept, które zostały przepisane nieprawidłowo (brak ZdUpr) po zamodelowaniu OWL, wymaga wyłącznie zdefiniowania zapytania typu SELECT języka SPARQL.



Rys. 4. Struktura właściwości systemu ControlSem

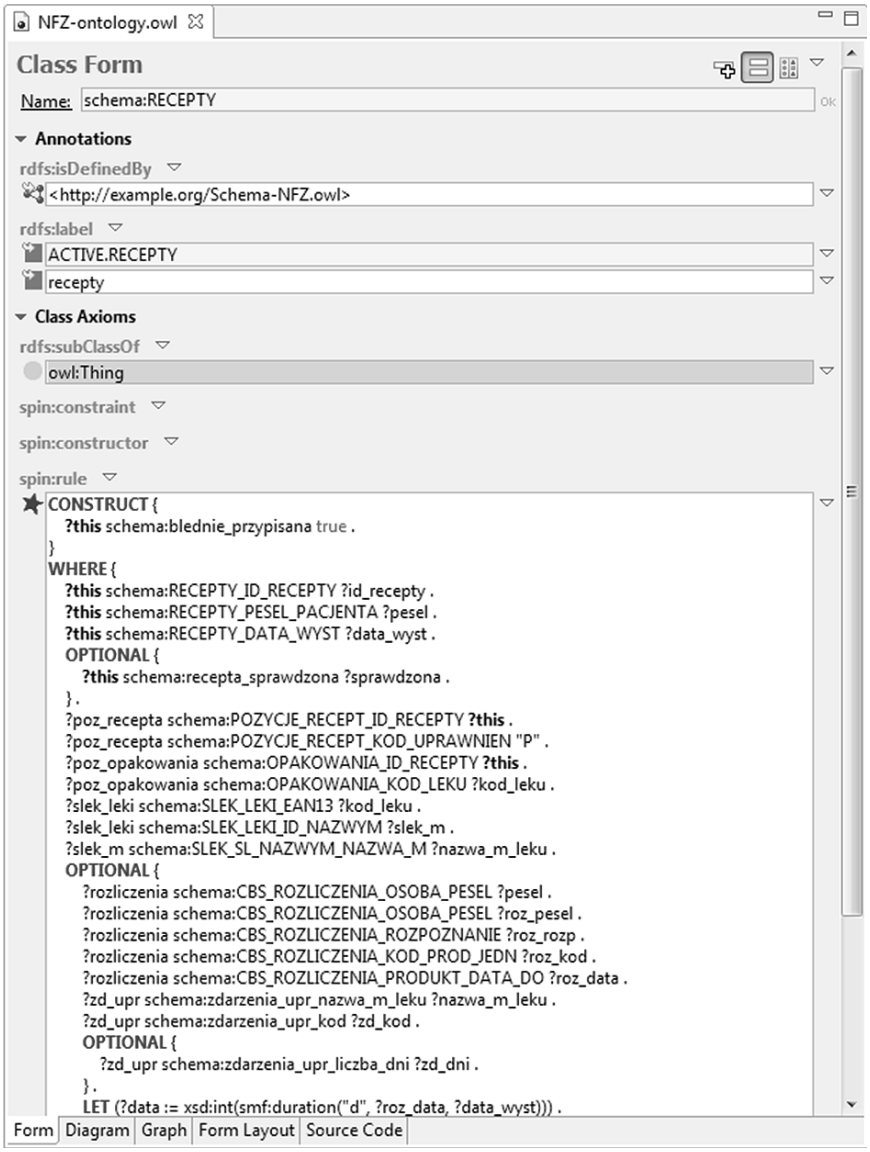


Rys. 5. Zapytanie SPARQL i identyfikatory recept – wynik działania zapytania

Dla przypadku (ii) zapytanie w SPARQL musiałoby uwzględniać tyle dodatkowych warunków, ile jest różnych okresów odpowiadających ZdUpr w słowniku zdarzeń. Dla tej grupy problemów przydatne jest zastosowanie zapytania typu Construct języka SPARQL w definicji reguły. W zapisie reguły uwzględniono wszystkie warunki definiowane przez lekarzy kontrolerów.

Najważniejsze cechy SPIN to:
- obliczanie wartości właściwości w oparciu o inne właściwości,
- sprawdzanie ograniczeń i sprawdzanie poprawności danych,
- tworzenie szablonów reguł wykonywanych przy spełnieniu określonych warunków.

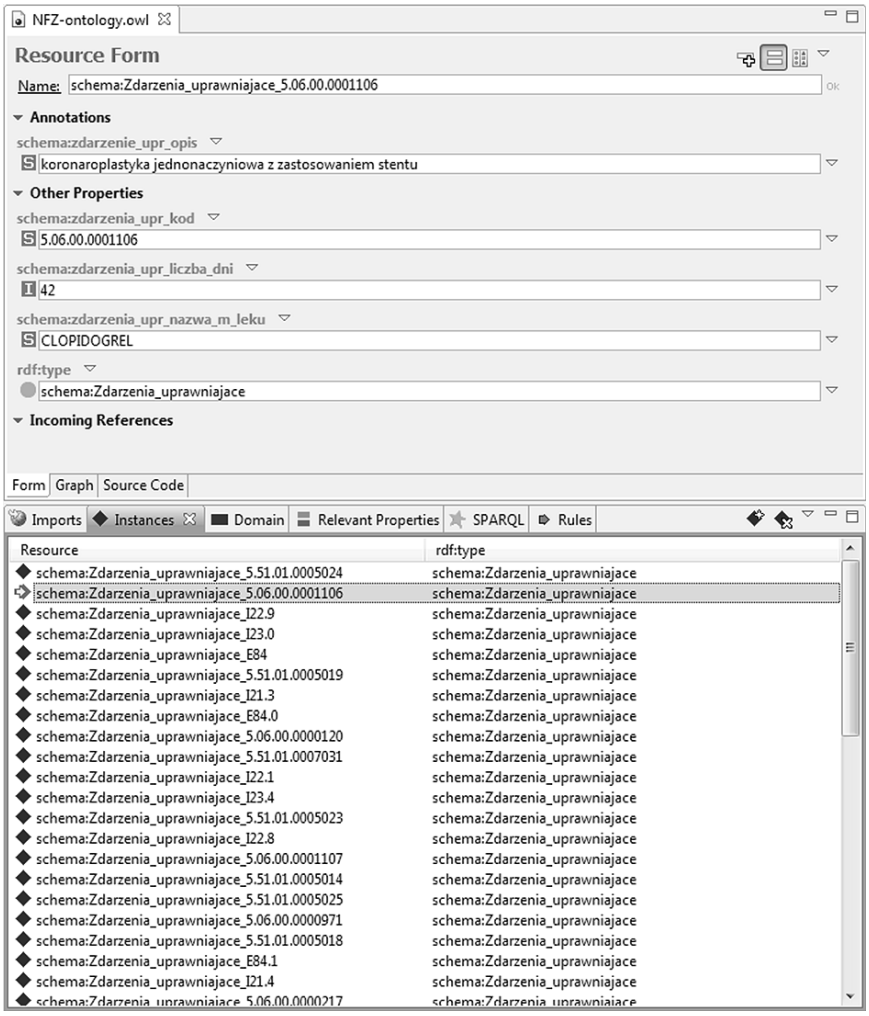
Na rys. 6 przedstawiono regułę zdefiniowaną w SPIN. Struktura tej reguły pozwala na tworzenie biblioteki podobnych zapytań, co stanowić będzie znaczące ułatwienie dla lekarzy kontrolerów.



Rys. 6. Reguła SPIN definiowana w TBC

W celu analizy przypadku (ii) wprowadzono do diagramu klas klasę Zdarzenia_uprawniajace z właściwościami: Zdarzenie_upr, Zdarzenie_upr_kod, Zdarzenie_upr_liczba_dni, Zdarzenie_upr_nazwa_m_leku typu *data properties*, które umożliwiają odwołanie się do wartości słowników ICD 9 i ICD 10.

W modelu ControlSem wsparcie ze strony TBC i wbudowanych SPARQL i SPIN pozwala na tworzenie optymalnych struktur zapytań, ale też umożliwia dalszy, elastyczny rozwój modelu systemu o kolejne struktury, reguły i zapytania. Ponadto, zastosowanie TBC pozwala na opracowanie przyjaznych interfejsów dla użytkowników końcowych systemu – lekarzy specjalistów.



Rys. 7. Instancje klasy Zdarzenia_uprawniajace z definiowana w TBC

4. Efektywność czasowa

Dla potwierdzenia poprawności generowanych przez ControlSem wyników została przygotowana baza zawierająca dane recept o znanej liczbie i identyfikatorach recept wystawionych nieprawidłowo. Po konwersji danych relacyjnych z wykorzystaniem D2RQ testowy zbiór danych składał się z 3 452 958 trójek. W wyniku zapytań SPARQL (przypadek i) oraz reguły SPIN (przypadek ii) otrzymano listy identyfikatorów recept wykazujących niezgodności z regułami ich poprawności.

Po stwierdzeniu zgodności wygenerowanych przez ControlSem wyników z danymi recept zgromadzonymi w bazie relacyjnej zostały przeprowadzone testy efektywności czasowej.

Parametry techniczne środowiska testowego – serwera wykorzystanego do zamodelowania architektury ControlSem:

1. Serwer:
   - procesor Intel Core 2 Duo E7600,
   - pamięć operacyjna 8 GB;
2. Oprogramowanie:
   - system operacyjny Windows Server 2003 R2,
   - serwer baz danych Oracle 10.2.0.1.0,
   - serwer baz danych Oracle 11.1.0.6.0,
   - oprogramowanie modelowania sieci semantycznych TopBraid Composer w wersji ME 3.2.0.

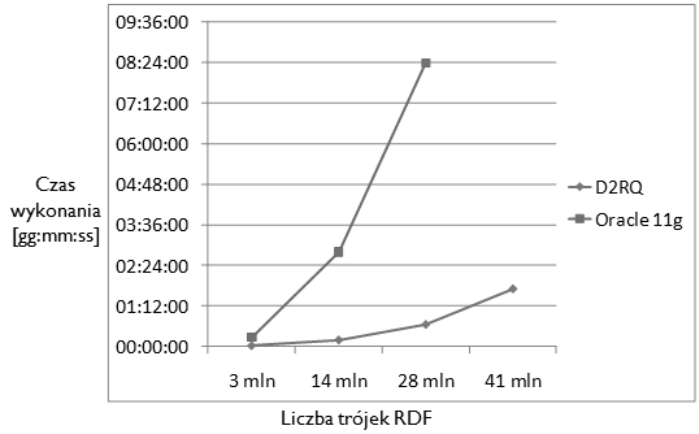
Tab. 1 oraz rys. 8 przedstawiają czasy wykonania analiz przy wykorzystaniu Oracle 11g jako bazy danych trójek RDF oraz narzędzia D2RQ współpracującego z relacyjną bazą danych i z uwzględnieniem różnej liczby trójek poddawanych analizom. Dla uzyskania większej liczby trójek w kolejnych zbiorach testowych (X5, X10, X15) dane wejściowe zostały powielone odpowiednio pięć, dziesięć i piętnaście razy poprzez kilkukrotne złączenie testowego zbioru danych (X1).

Wielkość zbioru danych:

1. dane x1 otrzymane od NFZ (3 452 958 trójek),
2. dane x5 bez tabel słownikowych leków (14 233 326 trójek),
3. dane x10 bez tabel słownikowych leków (27 708 786 trójek),
4. dane x15 bez tabel słownikowych leków (41 184 246 trójek).

Tab. 1. Czasy wykonania analiz dla różnych konfiguracji ControlSem

Lp.	Baza danych	D2RQ (gg:mm:ss)	Oracle 11g (gg:mm:ss)
1	X_1	00:01:09	00:16:00
2	X_5	00:11:10	02:47:30
3	X_{10}	00:38:54	08:23:30
4	X_{15}	01:42:10	>24:00:00



Rys. 8. Czasy wykonania analiz dla różnej liczby trójek RDF i różnych konfiguracji ControlSem

W warunkach dużej ilości przetwarzanych danych, co ma miejsce w przypadku zasobów NFZ, wydajność narzędzi semantycznych nie jest satysfakcjonująca dla analiz realizowanych w czasie rzeczywistym. Jednakże w porównaniu do narzędzi sposobu realizacji analiz w hurtowniach danych analityki biznesowej (*Business Intelligence*), gdzie zapytania są realizowane w okresach sprawozdawczych (np. raz na miesiąc czy raz na kwartał), można uznać, że reżim czasowy przetwarzania zarówno dla sieci semantycznych, jak i narzędzi analityki biznesowej jest porównywalny.

W celu optymalnego doboru istniejących technologii oprócz wielu testów przeprowadzonych na relacyjnej bazie danych (Oracle 10g) NFZ wykonano również szereg testów z wykorzystaniem semantycznego silnika Oracle 11g. Testy te wykazały, że wydajność Oracle 11g dla sieci semantycznych jest ok. 15 razy gorsza w porównaniu z połączeniem Oracle 9i oraz D2RQ.

Wnioski

Wynikiem eksperymentu jest osiągnięcie wysokiego stopnia integracji danych celem uzyskania oczekiwanych raportów bez konieczności jakiejkolwiek ingerencji w dotychczasową, rozproszoną strukturę baz danych.

Przyjęty model ControlSem oparty na narzędziach TBC oraz D2RQ potwierdził przydatność SW w analizie procedur medycznych opisanych rekordami zapisanymi w relacyjnych bazach danych.

Doświadczenia z modelowania ControlSem mogą być wykorzystane do:

- rozbudowy przyjętego modelu o kolejne struktury danych, zależności oraz reguły wnioskowania i narzędzia wizualizacji wyników danych,
- współdzielenia wiedzy zapisanej w opracowanych strukturach RDF, RDFS, OWL i regułach SPIN/SPARQLMotion w opisie procedur medycznych,
- wdrażania narzędzi SW opartych na podobnym modelu danych (np. duże bazy danych przedsiębiorstw – przychody, rozchody, wartości wskaźnikowe) lub analogiczne cele wnioskowania (np. wyszukiwanie nieprawidłowości w procesach biznesowych przedsiębiorstw, wielokryterialna klasyfikacja podmiotów gospodarczych czy wsparcie w ocenie szeroko rozumianej kondycji przedsiębiorstwa),
- implementacji rozwiązań kontekstowej analizy danych.

Literatura

Błaszczyński J., Kosiedowski M., Mazurek C., Wilk S. (2006): *Ontologies for Knowledge Modeling and Creating User Interface in the Framework of Telemedical Portal*, European Conference on eHealth, Fribourgh, Switzerland, Proceedings of the ECEH, s. 275–286.

Casteleiro M.A., Des Diz J.J. (2008): *Clinical Practice Guidelines: A Case Study of Combining OWL-S, OWL, and SWRL*, „Knowledge-Based Systems", vol. 21, s. 247–255.

Chen H., Wang Y., Wang H., Mao Y., Tang J., Zhou C., Yin A., Wu Z. (2006): *Towards a Semantic Web of Relational Databases: a Practical Semantic Toolkit and an In-Use Case from Traditional Chinese Medicine*, „Fifth International Semantic Web Conference", vol. 4273, http://iswc2006.semanticweb.org/items/Chen2006kx.pdf, s. 750–763.

Hebeler J., Fisher M., Blace R., Perez-Lopez A. (2009): *Semantic Web Programming*, Wiley Publishing, Inc., Indianapolis.

Kataria P., Juric R., Paurobally S., Madani K. (2008): *Implementation of Ontology for Inteligent Hospital Wards*, Proceedings of the 41st Hawaii International Conference on System Sciences (HICSS–2008).

LePendu P., Dou D., Frishkoff G.A., Rong J. (2008): *Ontology Database: a New IVlethod for Semantic Modeling and an Application to Brainwave Data*, [w:] *International Conference on Statistical and Scientific Database Alanagement*, s. 313–330.

Mirhaji P., Casscells S.W., Allemang D., Coyne R. (2007): *Improving the Public Health Information Network through Semantic Modeling*, „IEEE Intelligent Systems", May/June.

Oracle Database 11g Semantic Technologies. Semantic Data Integration for the Enterprise (2009), „An Oracle White Paper", September.

Rector A.L., Qamar R., Marley T. (2009): *Binding Ontologies and Coding Systems to Electronic Health Records and Messages*, „Applied Ontology", vol. 4, IOS Press, s. 51–69.

Semantic Technologies Software – Documentation,
http://www.oracle.com/technology/tech/semantic_technologies/pdf/semantic11g_dataint_twp.pdf.

The D2RQ Plattform – Treating Non-RDF Databases as Virtual RDF Graphs, http://www4.wiwiss.fu-berlin.de/bizer/d2rq/.