Encyklopedia semantyczna – odpowiedź na wymagania dietetyka

Semantic encyclopedia – a response towards requirements of dieticians

Krzysztof Sarapata [1/], Paweł Góra [1/], Grzegorz J. Nalepa [2/], Weronika T. Adrian [2/], Krzysztof Kluza [2/], Marta Noga [2/]

[1/] Zakład Bioinformatyki i Telemedycyny, Collegium Medicum Uniwersytet Jagielloński, Kraków
[2/] Katedra Automatyki, Akademia Górniczo-Hutnicza, Kraków

Projekt encyklopedii dietetyczno-medycznej w wersjach: eHealthWiki i Diki (Diet Wiki) jest przykładem internetowej encyklopedii typu wiki poszerzonej o moduł semantyczny. Tematyka encyklopedii plasuje się na styku medycyny i dietetyki. W pracy zaprezentowane zostały możliwości, które wprowadza encyklopedia semantyczna, konkretnie Semantic MediaWiki (SMW). Ponadto pokazano przykładowy rekord wraz z zależnościami semantycznymi dotyczący diety eliminacyjnej zalecanej w celiakii. Warto zwrócić uwagę, że oprócz waloru centralizacji i aktualności informacji zawartej w encyklopedii, projekt posiada aspiracje naukowe, jako narzędzie metodologiczne w dietetyce. Wykorzystanie zasobów kategoryzacji dostępnych w biologii i medycynie (taksonomia biologiczna, baza toksyczności i inne) pozwala na grupowanie, własną kategoryzację i stawiania hipotez. W pracy pokazano to na przykładzie występowania glutenu w produktach pochodzących z tej samej gałęzi taksonomicznej czy występujących w grupie produktów pochodnych.

Słowa kluczowe: *dieta, żywienie człowieka, semantyka*

The project of dietetic and medical encyclopedia (eHealthWiki and Diki (Diet Wiki)) is a case of an Internet-based encyclopedia using a wiki system enhanced with a semantic module. The contents of the encyclopedia are related to both medicine and dietetics. In the paper new features provided by the Semantic MediaWiki system used in the encyclopedia are outlined. A use case of a diet for patients with celiac disease is presented. The objectives of the project are both scientific and methodological in the dietetic science. The encyclopedia uses widely available medical and biological databases to categorize and group concepts. The use case described in the paper concerns the existence of gluten in dietary products in the same taxonomical branch.

Key words: *diet, human nutrition, semantics*

© **Probl Hig Epidemiol 2010, 91(4): 517-521**

www.phie.pl

Nadesłano: 10.08.2010
Zakwalifikowano do druku: 27.11.2010

Adres do korespondencji / Address for correspondence
Mgr inż. Krzysztof Sarapata
Zakład Bioinformatyki i Telemedycyny, Uniwersytet Jagielloński
Collegium Medicum, ul. Kopernika 7, 31-501 Kraków
tel./fax 12-422-77-64, e-mail: mysarapa@cyf-kr.edu.pl

Wstęp

Dostępne oprogramowanie wspierające pracę dietetyków: Wikt [1], Programy Dietetyk i Dieta 4 [2], Dietetyk [3], Energia [4] ma zastosowanie w dziedzinie planowania i analizy żywienia człowieka. Niezależnie od praktycznego stosowania programów informatycznych zarówno lekarz jak i dietetyk zobowiązani są do regularnego poszerzania wiedzy na podstawie nowych doniesień naukowych, a w konsekwencji do aktualizacji receptur i diet w różnego rodzaju chorobach. Encyklopedia eHealthWiki poprzez swoje walory staje się użytecznym narzędziem ułatwiającym zarówno aktualizację wiedzy, jak i wspomaganie leczenia w dziedzinie żywieniowej.

Powstająca internetowa encyklopedia eHealthWiki zbiera informacje dostępne w źródłach internetowych i papierowych dotyczących medycyny oraz problematyki żywienia człowieka. Jej celem jest koncentracja, ale nie powielanie, wiedzy o tematyce żywieniowej i medycznej. Wykorzystuje zasoby istniejące w Internecie oraz w wersji drukowanej, które sukcesywnie przenosi się do postaci elektronicznej. Dzięki wprowadzeniu oryginalnej struktury przechowywanych danych, encyklopedia umożliwia dynamiczne kategoryzowanie wg kryterium najistotniejszego z punktu widzenia żywienia człowieka określania stopnia toksyczności produktów żywnościowych. Encyklopedia bazuje na wypracowanych koncepcjach w dietetyce [5,6,7].

Dostępne systemy kategoryzowania danych biologicznych i medycznych, np. taksonomia biologiczna stosują różne podejścia i kryteria grupowania. Przykładowo tzw. analizę filogenetyczną przeprowadza się na podstawie sekwencji genomowych rodów organizmów i opiera się na określaniu ilości kroków potrzebnych

do przejścia z jednej do drugiej sekwencji nukleotydowej. Taksonomia oparta na strukturach białkowych wykorzystuje ocenę stopnia złożoności przestrzennej struktury. W domenie dietetyki zasadnym wydaje się wykorzystanie istniejących struktur danych do utworzenia własnej kategoryzacji według kryterium „wyższego" poziomu – wpływu substancji na organizm ludzki. Taka ocena wymaga wiedzy i danych klinicznych połączonych z wiedzą podstawową: od biologii i mechanizmów biochemicznych, poprzez odżywianie, aż do bromatologii. Pełna ocena całościowego wpływu jest jednak złożonym procesem, ponieważ wymaga uwzględnienia czynników czasowo-przestrzennych, historycznych, genetycznych, rodzinnych, interakcji między substancjami, itd. Ta wielorakość czynników stanowi uzasadnienie dla powstawania różnych systemów opierających się na odmiennych kryteriach kategoryzacji. Systemy te mogą osiągać specjalizację w wąskiej dziedzinie.

Specyfika problematyki żywienia człowieka staje się inspiracją rozwoju i propagowania encyklopedii. Główne cele encyklopedii to:

1. Koncentracja wiedzy dietetycznej i okołodietetycznej
2. Automatyczna aktualizacja danych na stronach receptur i diet
3. Reakcja na tzw. sprzeczność „metodologiczno-praktyczną" w relacji lekarz-pacjent.

Sprzeczność „metodologiczno-praktyczna" w relacji lekarz-pacjent rozumiemy jako obiektywną potrzebę indywidualnego traktowania pacjenta. Z kolei obiektywizm w metodach naukowych wymaga kolektywizmu. Pacjent oczekuje indywidualnego potraktowania przez lekarza, który opiera się jednak na metodologii medycyny zakładającej także ilościowe potwierdzenia hipotezy. Każda systematyka pozwala na doprowadzenie informacji od ogólności do szczegółowości, wciąż teoretycznie, ale stanowi to zdaniem autorów argument na rzecz rozwijania proponowanego systemu.

Systemy semantyczne

Systemy semantyczne pozwalają na taką reprezentację wiedzy, aby znaczenie danych było „zrozumiałe" dla komputerów i możliwe do automatycznego przetwarzania. Za pomocą odpowiednich mechanizmów systemy semantyczne mogą zagwarantować znaczące poszerzenie możliwości reprezentacji wiedzy biologiczno-molekularnej. Technologie i języki semantyczne: *Resource Description Framework* (RDF) [8], *RDF Schema* (RDFS) [9], czy *Web Ontology Language* (OWL) [10] w różnym stopniu pozwalają na reprezentację zdań z języka naturalnego, który stanowi oryginalny język publikacji naukowych. Ubogość języka na poziomie opisu mechanizmów biochemicznych pozwala na modelowanie relacji pomiędzy produktami genów i samymi genami już na poziomie modelu RDF, w którym stwierdzenia mają postać tzw. trójek: <subject> <predicate> <object>. Odpowiadają one sformułowaniom typu: obiekt <subject> ma własność <predicate> o wartości <object> lub: zasób <subject> pozostaje w relacji <predicate> z zasobem <object>.

Przykład systemu wykorzystującego model RDF to projekt *Gene Ontology* – GO [11]. W projekcie wyróżnia się trzy podgrupy stanowiące oddzielne obszary badań:

- kryterium funkcji molekularnej (*MF-molecular functions*),
- udział w procesach biologicznych (*BP-biological processes*),
- miejsce występowania w składnikach komórki (*CC-cellular components*)

W zależności od obszaru badań dany produkt genowy może być różnie scharakteryzowany, np. cytochrom c może być opisany jako:

- oksydoreduktaza (ze względu na funkcje molekularne),
- fosforylator oksydacyjny i induktor apoptozy (ze względu na udział w procesach biologicznych),
- składnik matriks mitochondrium i błony wewnętrznej mitochondrium (ze względu na występowanie w komórce).

Przykładowe terminy użyte jako własności – predicate w ontologii GO:

term_id	name
18099	PART_OF
18100	IS_A
18101	NEGATIVELY_REGULATES
18102	REGULATES
18103	POSITIVELY_REGULATES

Przykład zdania z GO:

ts	tp	name
negative regulation of nurse cell apoptosis	NEGATIVELY_REGULATES	nurse cell apoptosis

Technologie semantyczne znalazły praktyczne zastosowanie w systemach semantycznych wiki. Najpopularniejszą implementacją jest MediaWiki z rozszerzeniem SemanticMedia Wiki [12]. SemanticMediaWiki pozwala na utworzenie struktury bazy wiedzy poprzez wykorzystanie adnotacji semantycznych, definiowanie kategorii, własności i atrybutów pojęć występujących w encyklopedii. System obsługuje język zapytań ASK, który pozwala na dynamiczne generowanie listy pojęć spełniających zadane kryteria. Eksport danych do formatu RDF umożliwia stosowanie zewnętrznych narzędzi do jeszcze lepszego wykorzystania wiedzy zgromadzonej w wiki. Odpowiednie mechanizmy wnioskowania stosowane w semantycznych wiki pozwalają na wykorzystanie wiedzy zgromadzonej w systemie do uzyskiwania dodatkowych informacji [13,14]. Wnioskowanie

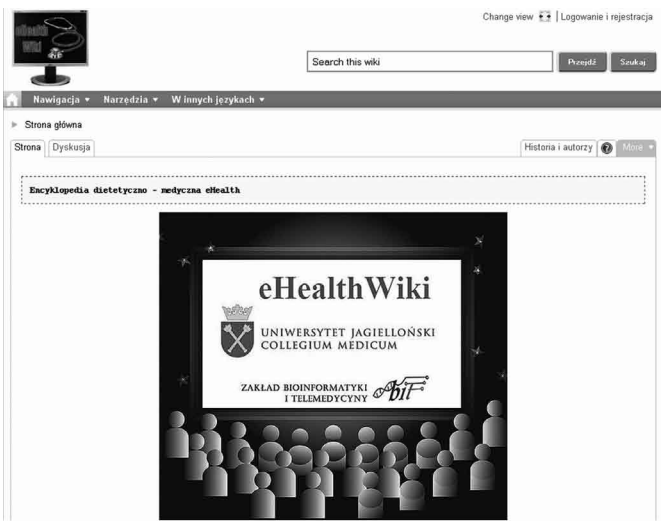
w Semantic MediaWiki wykorzystuje formalizm Description Logic (DL) [15]: zapytania do bazy DL opierają się na tworzeniu opisów pojęć. Pojawiające się ograniczenia to przede wszystkim brak zmiennych oraz możliwości definiowania reguł. Istniejące implementacje semantycznych wiki wspierających reguły to np. PlWiki [16] czy KnowWE [17].

Praktyczne możliwości SMW wykorzystane w projekcie encyklopedii eHealthWiki to:

- Automatycznie generowane listy pojęć spełniających zadane kryterium,
- Wizualizacja informacji w różnych formatach,
- Elestyczna struktura danych, możliwość generowania dynamicznych kategorii
- Zewnętrzna możliwość wykorzystania danych,
- Mechanizm wyszukiwania informacji: rozszerzenia „Halo" i „Semantic Drilldown"

Projekt encyklopedii dietetycznej

Rozwój oprogramowania wspierającego prace specjalistów zarówno dietetyków jak i lekarzy zmierza w kierunku automatycznego przetwarzania danych, w taki sposób, aby ułatwić określenie związków pomiędzy dietą a chorobą. Proponowany kierunek ma za cel oznaczenie i definiowane relacji, tworzenie sieci zależności między już istniejącymi zasobami czy to w postaci zorganizowanej (drzewa, grafy), np. kategoryzacja produktów spożywczych wg Kunachowicz [7], baza taksonomii biologicznej wg NCBI [18], czy tylko w postaci „prostych" zależności produktu żywieniowego z chorobą (np. baza TOXNET [19]). Istotnym staje się zdefiniowanie tych zależności z wykorzystaniem różnych metod eksploracyjnych. Mogą to być techniki statystyczne lub konkretne (*case study*), które posiadają znaczenie rozwojowe, naukowe.



Ryc. 1. Strona WWW encyklopedii eHealthWiki

Fig. 1. eHealthWiki encyclopedia website

Zdefiniowanie relacji między dietą a chorobą wymaga przygotowania systemu semantycznego zarówno w domenie dietetyki jak i patologii. W tym celu lokalnie zaimplementowano w oryginalnej strukturze zasoby kategoryzacji produktów spożywczych wg Kunachowicz, dodatków E, norm żywieniowych, taksonomie biologiczną, ICD [20] i inne. Zachowanie oryginalnej struktury tych zasobów ułatwia w przyszłości jej aktualizację. Następnie utworzono własną strukturę tych danych opartą na dynamicznych kategoriach. Rozwiązanie to jest dostępne na serwerze WWW Collegium Medicum UJ pod nazwą „Encyklopedia żywieniowo-medyczna E-Health" http://pl.ehealth.cm-uj.krakow.pl (ryc. 1).

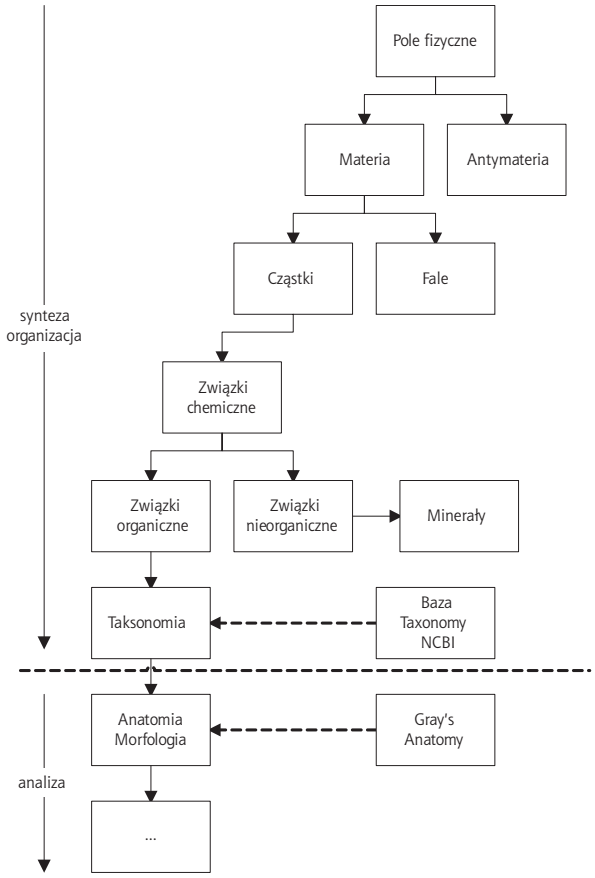
Zalety przedstawionego rozwiązania, w szczególności organizacji danych, można przedstawić na przykładzie jednostki chorobowej celiakia. Zaznaczając gluten jako składnik niepożądany w diecie, w jednym miejscu, automatycznie wykreślimy wszystkie produkty zawierające ten składnik lub produkty „podejrzane" tzn. takie które znajdują się w małej „odległości" toksycznej od glutenu. Inny przykład dotyczy skazy białkowej: zaznaczając mleko, domyślnie wyłączamy z diety wszystko, co jest „poniżej" w drzewie kategoryzacji. Powiązanie diety z różnego rodzaju kategoryzacjami, taksonomiami, służy do koncentracji wypracowanych schematów informacji w jednym miejscu dla konkretnego celu.

W encyklopedii wprowadzono kategorię główną oraz kategorie pomocnicze. Kategoria główna powstaje przez poszerzenie struktury taksonomii biologicznej zarówno „w dół" – metody analityczne i jak i „w górę" – metody syntetyczne. Zostało to przedstawione na schemacie blokowym (ryc. 2).

Jako kategorie pomocnicze wykorzystano (ryc. 3):

- Bazę produktów żywnościowych: Baza Kunachowicz,
- Bazę toksykologiczną Toxnet (Toksykologiczna baza danych udostępniona przez National Library of Medicine),
- Bazę chorób ICD9/10 [20],
- Bazę związków chemicznych [21,22],
- Diety,
- Rozwój osobniczy (embrionalny, dojrzewanie,...),
- Proces kulinarny.

Referencje bibliograficzne wprowadzanych danych czy zależności przechowywane są w plikach o formacie bibtex [23]. Jest to format uniwersalny, dla którego istnieją gotowe rozwiązania wspierające import i eksport z MediaWiki. Z Semantic MediaWiki można eksportować (Semantic Result Formats)[24] i importować (Biblio Extension)[25] informacje bibliograficzne w formacie bibtex. Bibtex dopuszcza



Ryc. 2. Główne drzewo kategoryzacji

Fig. 2. Main categorization tree

także tworzenie własnych tagów, zatem można wprowadzić dodatkowe pole np. kategoria, aby uwzględnić tematykę. Możliwe jest także grupowanie i przeglądanie bibliografii wg zadanych kryteriów.

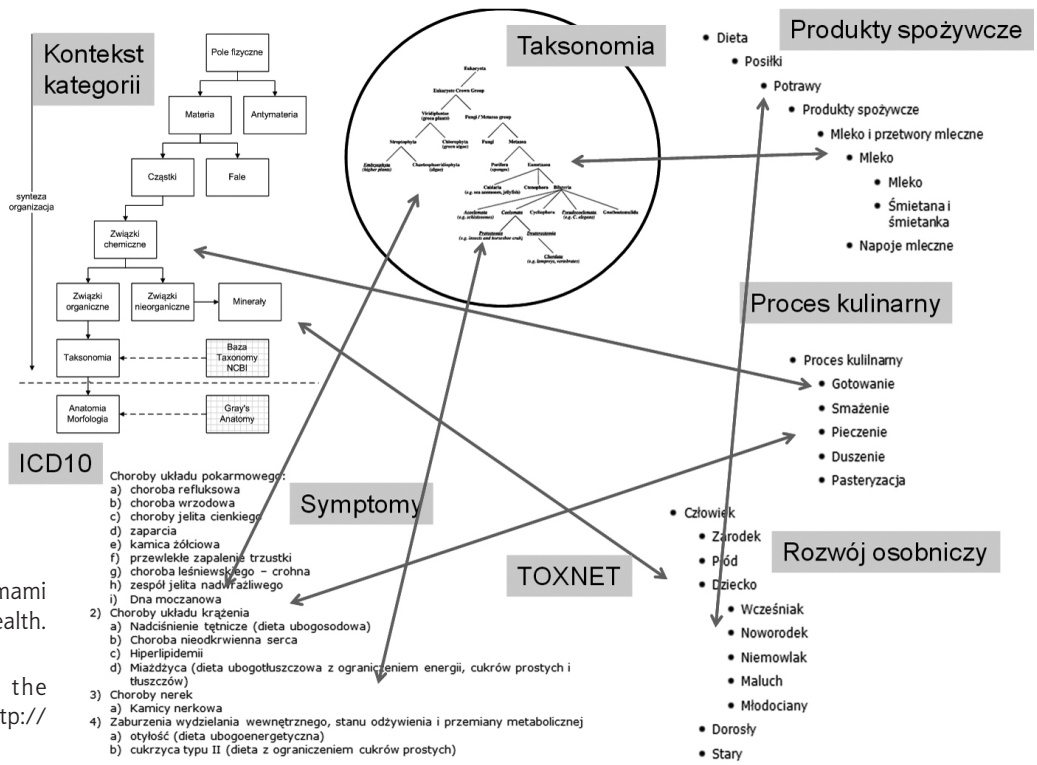
Przypadek użycia

Schemat relacji „choroba-dieta" został przedstawiony na przykładzie zależności: celiakia – gluten [26], wprowadzony do założonej struktury danych w encyklopedii.

Wykorzystanie kategoryzacji w obrębie taksonomii biologicznej stanowi element metodologii określania i przewidywania toksyczności produktów żywnościowych pochodzących z danej gałęzi struktury hierarchicznej. Tutaj rodzaj „gryka" pochodzący z innego taksonu wyraźnie sugeruje nietoksyczność dla osób z rozpoznaną celiakią (ryc. 4).

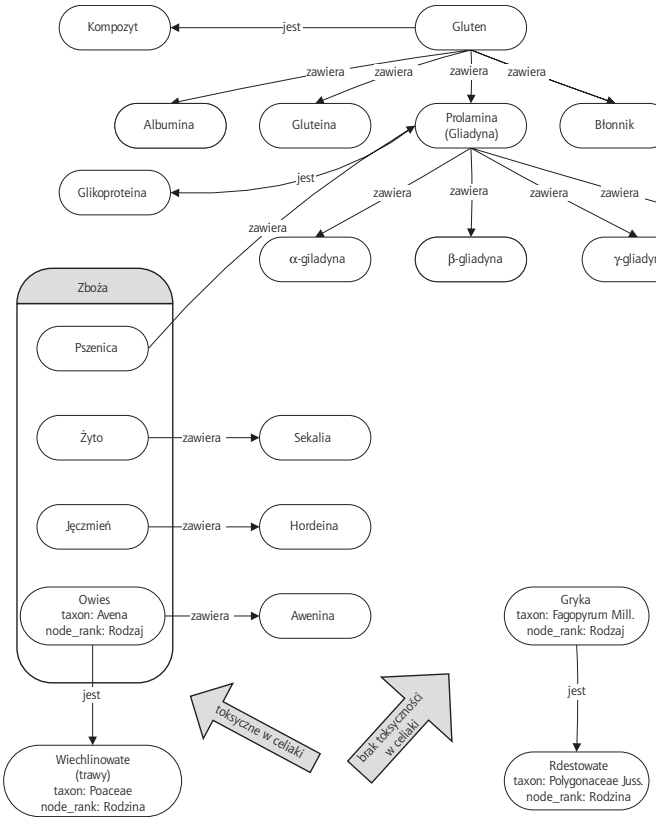
Zakłada się, że poszerzenie schematu struktury danych jeszcze bardziej poprawi przewidywalność i wnioskowanie w dziedzinie dietetyki.

Struktura danych zostaje w taki sposób zaprojektowana, aby można było wygenerować dynamiczną kategoryzację bazującą na taksonomii biologicznej ale z uwzględnieniem kryterium toksyczności.



Ryc. 3. Zależności między systemami klasyfikacji (źródło: http://pl.ehealth.cm-uj.krakow.pl)

Fig. 3. Correlations among the classification systems (source: http://pl.ehealth.cm-uj.krakow.pl)



Ryc. 4. Fragment struktury danych dotyczący relacji glutenu

Fig. 4. Fragment of data structure on gluten relations

Podsumowanie

Rozwijanie encyklopedii internetowej typu wiki z możliwością tworzenia struktury semantycznej pozwala na:

- Kategoryzację produktów żywnościowych wg kryterium „szkodzi-pomaga" (toksyczność – ożywczość), z możliwością gradacji tej cechy
- Uwzględnienie kryterium „toksyczność – ożywczość" w taksonomii biologicznej
- Organizację danych umożliwiającą łatwą aktualizację wiedzy

Organizacja i koncentracja danych powinna także uwzględniać metody, którymi posługiwano się dokonując analizy składu produktów. Jest to związane z różnym poziomem opisu składu danej substancji lub zdolności danej metody do wykrywania śladowych związków występujących w danych produktach.

Piśmiennictwo / References

1. NET-KOMP Usługi informatyczne. Gdynia. http://www.wikt.pl
2. Zakład Epidemiologii i Norm Żywienia. Instytutu Żywności i Żywienia, Warszawa.
3. Alpha-NET Software. www.dietetyk.com.pl
4. http://www.energia.waw.pl
5. Willett WC, Stampfer MJ. Rebuilding the Food Pyramid. Scientific American 2003, 288(1): 64-71.
6. Holland B, Welch AA, Unwin ID, Buss DH, Paul AA, Southgate DAT. McCance and Widdowson's The Composition of Foods, Royal Society of Chemistry, Cambridge 1991.
7. Kunachowicz H, Nadolna I, Przygoda B, Iwanow K. Tabele składu i wartości odżywczej żywności. PZWL, Warszawa 2005.
8. http://www.w3.org/RDF/
9. http://www.w3.org/2001/sw/wiki/RDFS
10. http://www.w3.org/2001/sw/wiki/OWL
11. http://www.geneontology.org
12. http://semantic-mediawiki.org/
13. Krötzsch M, Schaffert S, Vrandečić D. Reasoning in semantic wikis. [in:] Antoniou G, et al (eds). 3rd Reasoning Web Summer School, volume 4636 of LNCS. Springer, 2007.
14. Noga M, Kaczor K, Nalepa GJ. Lightweight Reasoning Methods in Selected Semantic Wikis. Zesz Nauk ETI Politechniki Gdańskiej 2010, 18: 103-108.
15. http://dl.kr.org/
16. Nalepa GJ. PlWiki – a generic semantic wiki architecture. [in:] Nguyen NT, et al. (eds). 1st International Conference on Computational Collective Intelligence – Semantic Web, Social Networks & Multiagent Systems. Volume LNAI 5796, Springer 2009: 345-356.
17. Baumeister J, Reutelshoefer J, Puppe F, KnowWE. community-based knowledge capture with knowledge wikis. In K-CAP '07: Proceedings of the 4th international conference on Knowledge capture, NY, USA 2007: 189-190.
18. http://www.ncbi.nlm.nih.gov/Taxonomy
19. http://toxnet.nlm.nih.gov
20. http://www.who.int/classifications/icd
21. http://pubchem.ncbi.nlm.nih.gov/
22. CTD – Comparative Toxicogenomics Database: http://ctd.mdibl.org/
23. http://www.bibtex.org/
24. http://www.mediawiki.org/wiki/Extension:Semantic_Result_Formats/bibtex_format
25. http://www.mediawiki.org/wiki/Extension:Biblio
26. Wartość odżywcza produktów i potraw. Dieta bezglutenowa, co wybrać? Kunachowicz H (red). PZWL, Warszawa 2001.