Uniwersytet Warszawski

Wydział Matematyki, Informatyki i Mechaniki

Krzysztof Dudzik

Nr albumu: 248349

Aplikacja wspomagająca tworzenie i edycję haseł w polskim Wikisłowniku

Praca magisterska
na kierunku INFORMATYKA

Praca wykonana pod kierunkiem
dr. hab. Jerzego Tyszkiewicza, prof. UW
Instytut Informatyki

Czerwiec 2011

Oświadczenie kierującego pracą

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data Podpis kierującego pracą

Oświadczenie autora pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autora pracy

Streszczenie

Tematem pracy jest aplikacja służąca do ułatwienia pracy autorów haseł w polskim Wikisłowniku. Jej funkcje mają w maksymalny możliwy sposób ułatwić tworzenie i edytowanie haseł osobom bez wiedzy informatycznej i technicznej, a także automatyzować możliwie wiele rutynowych czynności wykonywanych przy redagowaniu hasła, jak tworzenie łącz do haseł powiązanych, zautomatyzowane szukanie przykładów użycia, wystąpień w związkach frazeologicznych, wyrazów bliskoznacznych, innych słów, którą formę gramatyczną mogłoby stanowić hasło itp. Dodatkowo aplikacja może przejąć część funkcji realizowanych obecnie za pomocą botów.

Słowa kluczowe

Wikisłownik, Fundacja Wikimedia, MediaWiki, wiki, edytor, API, JavaScript, jQuery, interfejs użytkownika, społeczność internetowa

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.3 Informatyka

Klasyfikacja tematyczna

D. Software
D.2. Software Engineering
D.2.10. Design

Tytuł pracy w języku angielskim

An application supporting article creation and edition for the Polish Wiktionary

Spis treści

Rozdział 1

Wprowadzenie

Żyjemy w czasach, w których nieustannie zmienia się sposób wyszukiwania informacji przez przeciętnego człowieka. Z roku na rok coraz mniejszą rolę odgrywają papierowe kompendia takie jak encyklopedie i słowniki, stopniowo przybierają natomiast na znaczeniu elektroniczne bazy wiedzy – szczególnie zaś internetowe zbiory danych. Przyczyny tego stanu rzeczy są oczywiste: chodzi przede wszystkim o wygodę korzystania ze stron internetowych. Brak możliwości wyszukiwania w obrębie ogromnych ilości danych powoduje, że encyklopedie i słowniki w postaci książek stają się o wiele mniej atrakcyjne dla kogoś, kto chce zdobyć nowe informacje.

Wszechobecny dostęp do internetu sprawia, że to właśnie w sieci WWW powstają najbardziej popularne bazy ludzkie wiedzy. Nie ma chyba internautów, którzy nie korzystaliby, rzadziej lub częściej, z Wikipedii – internetowej encyklopedii pisanej przez ochotników. Właśnie fakt, że encyklopedia ta współtworzona jest przez amatorów, stanowi o jej wyjątkowym charakterze, który zostanie w tej pracy pokrótce opisany. Wikipedia stale utrzymuje się w pierwszej dziesiątce najczęściej odwiedzanych stron, a pod wieloma względami jest to dziś najlepsza istniejąca encyklopedia. Przed kilkoma laty głośne było porównanie jej z prestiżową *Encyclopaedia Britannica* – okazało się, że różnice w poziomie merytorycznym są niewielkie.

O ile przewrót w kategorii encyklopedii właściwie już się dokonał, nieco inaczej wygląda rywalizacja słowników. Oczywiście wyraźnie widać, że i tu papierowe edycje są coraz mniej popularne. Różnice uwidaczniają się, gdy przeanalizowana zostanie sytuacja słowników internetowych. Tak zwany siostrzany projekt Wikipedii, Wikisłownik, nie dominuje wśród konkurencji – zarówno na świecie, jak i w Polsce. Przyczyny tego stanu rzeczy są złożone. Autor postanowił skupić się na kilku zagadnieniach, uwidaczniających się w polskojęzycznej wersji Wikisłownika. W tym celu konieczne było zbadanie społeczności zaangażowanej w tworzenie tego projektu. Jego efektem było wykonanie prac programistycznych, których opis stanowi główną część niniejszego opracowania.

W przypadku wszystkich projektów opartych na silniku programistycznym MediaWiki istotną barierą rozwoju jest sama technologia. Każdy ochotnik ma możliwość uczestniczenia w rozwoju portalu, wiąże się to jednak z koniecznością przystosowania się do wymagań stawianych przez oprogramowanie. Edytowanie haseł w internetowej encyklopedii czy słowniku jest praktycznie niemożliwe dla osoby bez wcześniejszego przygotowania lub znacznej wiedzy techniczno-informatycznej. Oprogramowanie MediaWiki oparte jest bowiem na tzw. wikikodzie (także: wikitekst, wikiskładnia), czyli języku opisu struktury i wyglądu strony internetowej – prostszym niż HTML, jednak wciąż nieintuicyjnym dla kogoś, kto nie miał wcześniej do czynienia z tego typu edytorami. Dlatego wielu potencjalnych współautorów zniechęca się do projektu już przy pierwszej próbie poprawy artykułu.

Aby zmienić tę sytuację, przygotowany został nowy edytor, dostosowany specjalnie do potrzeb polskiego Wikisłownika. Aplikacja pozwala na o wiele prostsze tworzenie nowych i zmienianie starych haseł niż poprzednia, standardowa. Dzięki użyciu jej jako domyślnej w projekcie popularyzacja edytowania Wikisłownika wśród fachowców w dziedzinach lingwistycznych okaże się łatwiejsze – zniknie podstawowa bariera, jaką jest konieczność dostosowania się do skomplikowanych technicznych wymagań stawianych przez użyte oprogramowanie. Dodatkowo nowa aplikacja umożliwia zaawansowaną automatyzację tworzenia hasła. Wiele z czynności zintegrowanych z nowym edytorem do tej pory wymagało mozolnych poszukiwań w artykułach Wikisłownika oraz innych projektach. Dzięki użyciu API udostępnianego przez serwisy Fundacji Wikimedia skomplikowane przeszukiwanie tysięcy stron udało się sprowadzić do kilku kliknięć.

W dalszej części pracy opisany został proces tworzenia tego edytora. Pierwszy rozdział charakteryzuje pokrótce sam Wikisłownik, jak i pokrewne projekty oraz oprogramowanie w nich użyte. Następnie opisano społecznościowe aspekty tworzenia tego typu aplikacji ze szczególnym uwzględnieniem koncepcji *wiki*. Ostatni rozdział wyczerpująco przedstawia szczegóły projektowe i implementacyjne aplikacji.

Rozdział 2

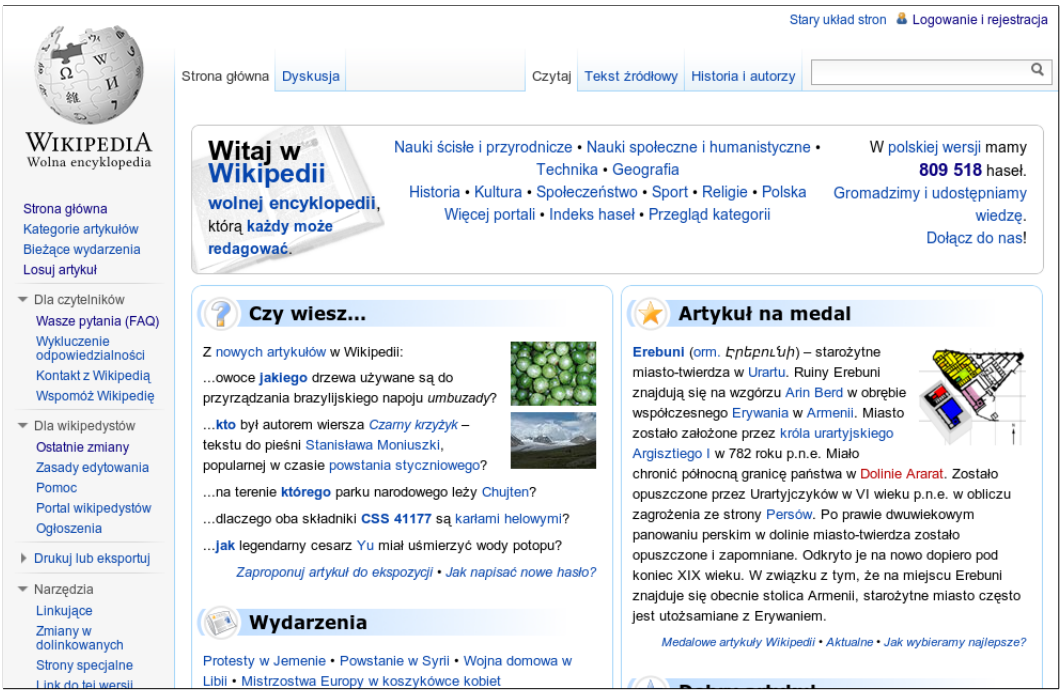
Wikisłownik

Rozdział ten stanowi charakterystykę Wikisłownika – sieciowego słownika opartego na oprogramowaniu MediaWiki. Wikisłownik jest jednym z największych i najpopularniejszych słowników dostępnych w polskim internecie. W kolejnych sekcjach projekt ten został opisany na różnych poziomach szczegółowości. Omówiono zarówno oprogramowanie, na jakim bazuje słownik, jak i swego rodzaju „ekosystem", w którym znajduje on swoje miejsce.

2.1. Projekty Fundacji Wikimedia

Podmiotem odpowiedzialnym m.in. za rozwój Wikisłownika jest Wikimedia Foundation Inc. (opisywana dalej jako „Fundacja") – organizacja non-profit mająca siedzibę w San Francisco w Stanach Zjednoczonych, istniejąca od 2003 roku. Jak informuje strona internetowa polskiego partnera Fundacji, Stowarzyszenia Wikimedia Polska, *celem fundacji jest sprzyjanie tworzeniu i rozwojowi projektów o otwartej treści opartych na technologii WikiWiki oraz dostarczanie społeczności internetowej pełnej zawartości wymienionych projektów za darmo i bez zamieszczania reklam.* Doskonale znaną, sztandarową inicjatywą Fundacji jest Wikipedia (`http://www.wikipedia.org`) – największa obecnie encyklopedia internetowa, dostępna w 281 językach (stan z maja 2011 roku) i zawierająca ponad 18 milionów haseł, w tym ponad 3,6 miliona w największej, angielskojęzycznej[1] edycji. Mimo częstej krytyki tego przedsięwzięcia faktem jest, że Wikipedia jest miejscem, z którego miliony osób korzystają, by pozyskać informacje z najróżniejszych dziedzin. Obecny stan rzeczy możliwy jest dzięki pracy wielkiej liczby wolontariuszy tworzących artykuły bez wynagrodzenia.

[1] Oficjalnie w projektach Fundacji używane są określenia typu *angielskojęzyczny*, *polskojęzyczny*. Choć w przypadku wersji polskojęzycznej znakomita większość uczestników projektów pochodzi z Polski, nie jest to regułą dla innych edycji. W dalszej części pracy przyjęto uproszczenie polegające na tym, że określenia typu *polska Wikipedia*, *angielski Wikisłownik* traktowane są jako tożsame z określeniami używającymi sformułowań z cząstką *-językowy*.



Ilustracja 2.1: Polska edycja Wikipedii

Wikipedia jest najbardziej znanym, ale nie jedynym projektem pod opieką Fundacji. Pozostałe to tzw. „projekty siostrzane", w szczególny sposób uwzględniane również przy tworzeniu haseł w encyklopedii. Oto lista wspieranych przez Fundację wielojęzycznych inicjatyw:

- Wikisłownik (ang. *Wiktionary*) – wielojęzyczny słownik internetowy, będący głównym przedmiotem niniejszej pracy,
- Wikicytaty (ang. *Wikiquotes*) – zbiór cytatów autorstwa znanych osób, z filmów i książek, przysłów i porzekadeł,
- Wikibooks – serwis z „otwartymi" (opartymi na wolnej licencji) podręcznikami,
- Wikiźródła (ang. *Wikisource*) – zbiór dokumentów źródłowych w wersjach oryginalnych i tłumaczonych, nieograniczonych prawem autorskim,
- Wikinews – otwarty serwis informacyjny,
- Wikimedia Commons – repozytorium mediów (zdjęć, grafik, filmów) dostępnych na wolnej licencji, z którego korzystają pozostałe projekty Wikimedia,
- Wikispecies – katalog gatunków organizmów żywych,
- Wikiversity – materiały edukacyjne i naukowe,
- Wikimedia Incubator – metaprojekt umożliwiający tworzenie nowych inicjatyw wspieranych przez Fundację.

- Meta-Wiki – projekt ułatwiający koordynację wszystkich pozostałych.

Wszystkie projekty łączy sposób ich powstawania – możliwość edycji dostępna jest praktycznie dla każdego internauty. Nie dotyczy to co prawda kilku krajów, w których projekty Fundacji zablokowane są w ramach cenzury internetu, jednak ogromna większość osób dysponujących łączem internetowym ma szansę stać się jednymi spośród współautorów haseł.

Drugą cechą wspólną są wolne licencje, na których udostępniana jest zawartość wszystkich serwisów. Po reformie w czerwcu 2009 roku treść Wikipedii i projektów siostrzanych dostępna jest nie tylko na licencji GNU FDL (Free Documentation License), ale także na kompatybilnej z nią CC-BY-SA 3.0 (Creative Commons Attribution-ShareAlike / Uznanie Autorstwa – Na Tych Samych Warunkach). Oznacza to, że można ją dowolnie wykorzystywać we własnych dziełach pod warunkiem podania oryginalnych autorów i zachowania pierwotnej licencji.

2.2. Oprogramowanie MediaWiki

Sama działalność wolontarystyczna redaktorów projektów Wikimedia nie wystarczyłaby do stworzenia serwisów internetowych o obecnych kształtach. Konieczne jest oczywiście również zapewnienie oprogramowania, które umożliwi płynną współpracę przy tworzeniu haseł. Tym oprogramowaniem jest wolna platforma MediaWiki tworzona zgodnie z zasadami *open source*. System MediaWiki napisany jest w języku PHP i obsługuje kilka popularnych baz danych (w przypadku projektów Wikimedia jest to MySQL). Dla inicjatyw Wikimedia stanowi szkielet programistyczny od samego ich początku, a od 2002 roku stale się rozwija. W czerwcu 2011 roku wersją używaną w projektach było MediaWiki 1.17.

System MediaWiki używany jest nie tylko w projektach wspieranych przez Fundację, ale także w tysiącach innych, mniejszych lub większych, co jest możliwe dzięki wysokiemu stopniowi konfigurowalności i dużej liczbie dostępnych rozszerzeń. Są to w dużej mierze serwisy o podobnym charakterze, umożliwiające swobodną wymianę informacji na dowolny temat. MediaWiki bywa także używane w firmowych intranetach i wszędzie tam, gdzie zachodzi potrzeba udostępnienia materiałów do edycji dużej liczbie użytkowników.

2.2.1. Edytowanie i wikitekst

Strony w projektach opartych na platformie MediaWiki na ogół nie mogą być czystym tekstem, pozbawionym formatowania. Przykładowo hasła w encyklopedii muszą zachowywać określoną strukturę – występuje więc podział na sekcje, ilustracje, różne rodzaje formatowania

(kursywa, wytłuszczenie), przypisy czy powtarzalne fragmenty. Szczególnie istotnym elementem są linki pomiędzy poszczególnymi artykułami, wyróżniające projekty Fundacji na tle ich papierowych, ale też elektronicznych konkurentów. Odnośniki pozwalają błyskawicznie przemieszczać się między hasłami, by w ten sposób uzyskiwać kolejne informacje wspomagające przyswajanie wiedzy.

Linki i formatowanie na stronach internetowych tworzone są za pomocą elementów języka HTML lub XHTML. O ile języki te są proste w obsłudze dla specjalisty informatyka, to laik nie jest w stanie tworzyć za ich pomocą stron bez uprzedniego dłuższego przygotowania. Aby umożliwić bezproblemową edycję stron internetowych osobom bez wykształcenia informatycznego, programiści MediaWiki zaprojektowali tzw. wikitekst – uproszczony język opisu stron, pozwalający na realizację wymienionych elementów. Porównanie niektórych z nich znajduje się w tabeli 2.2.1.

	Wikitekst	XHTML
Kursywa	''Tekst''	Tekst
Wytłuszczenie	'''Tekst'''	Tekst
Nagłówek	== Nagłówek == === Nagłówek ===	Nagłówek Nagłówek
Odnośnik wewnętrzny	[[Strona]] [[Strona|strony]]	Strona strony
Odnośnik zewnętrzny	[http://www.google.com Google]	 Google
Obraz	[[Plik:Przykład.png|thumb|Podpis]]	 Podpis
Podział na akapity	Pierwszy akapit Drugi akapit	Pierwszy akapit Drugi akapit
Lista nienumerowana	* Element * Element * Element	 - Element - Element - Element
Lista numerowana	# Element # Element # Element	 - Element - Element - Element

Tabela 2.1: Porównanie HTML i wikitekstu

Łatwo można zauważyć, że używanie wikitekstu jest o wiele prostsze niż nauka XHTML-a. Jeśli zachodzi potrzeba zaawansowanego formatowania, możliwe jest także użycie znaczników XHTML. W przypadku standardowego formatowania jest to jednak niewskazane ze względu na dobro niedoświadczonych edytorów.

Bardzo istotnym elementem wikitekstu są szablony – predefiniowane fragmenty kodu z opcjonalnymi parametrami. Szablony można uznać za odpowiednik procedur/funkcji w językach programowania. W projektach opartych na MediaWiki szablony pełnią przede wszyst-

kim dwie główne funkcje:

- upraszczają kod – pozwalają np. na zastąpienie skomplikowanego kodu XHTML (a także jeszcze bardziej złożonych funkcji parsera MediaWiki) krótkim wywołaniem szablonu,
- standaryzują strony – często wykorzystywane fragmenty wywoływane są zawsze w dokładnie ten sam sposób.

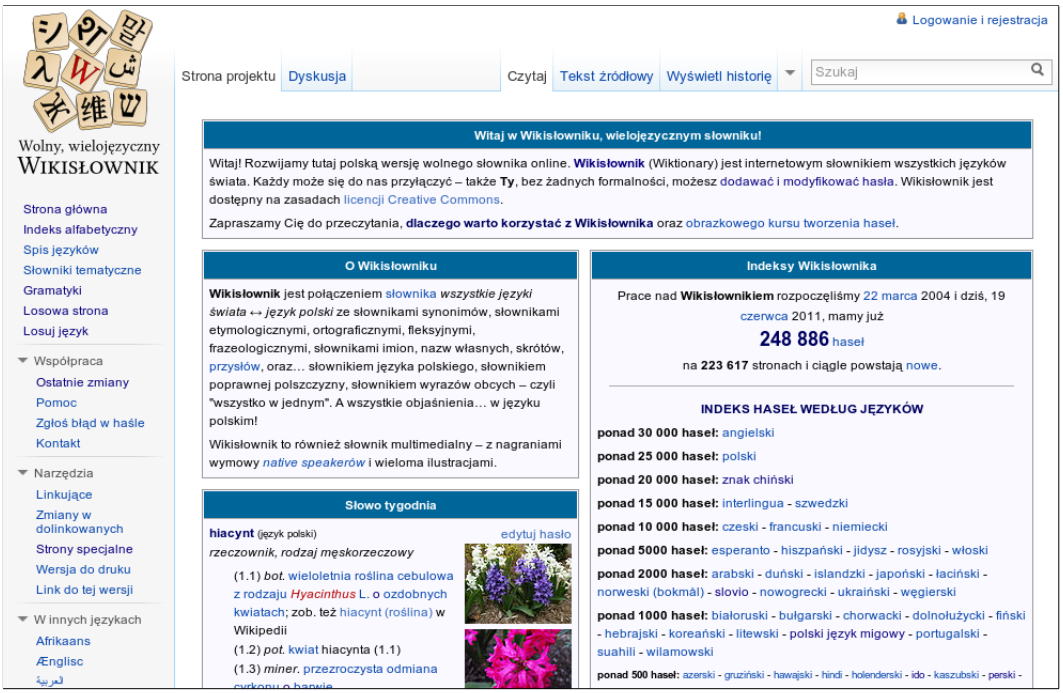
We wszystkich większych projektach Fundacji szablony są bardzo często wykorzystywane. Przy tym stanowią duże ułatwienie dla technicznie zaawansowanych autorów, którzy wspomagają się przy tworzeniu haseł dodatkowymi technologiami. Na używaniu szablonów korzystają przede wszystkim boty, czyli programy dokonujące edycji samodzielnie po uprzednim przygotowaniu lub pod stałą opieką programisty. W większości projektów przyjęte jest, że każdy bot ma własne konto użytkownika, nie używa natomiast konta swojego „właściciela". Dzięki botom możliwe jest np. masowe tworzenie haseł w Wikipedii na ściśle określony temat, jeśli istnieje dobre źródło w formie czytelnej dla komputera (takich jak opisy asteroid czy wszystkich miejscowości lub jednostek administracyjnych w danym kraju). Innym ich zastosowaniem jest automatycznie uzupełnianie tzw. interwiki – czyli odnośników pomiędzy poszczególnymi wersjami językowymi tego samego hasła.

Szablony mogą być wykorzystywane w prosty sposób nawet przez początkujących użytkowników – aby wywołać szablon, wystarczy wpisać jego nazwę pomiędzy podwójnymi nawiasami klamrowymi (`{{Nazwa szablonu}}` lub `{{Nazwa szablonu|parametr=wartość}}`). W polskim Wikisłowniku szablony pełnią szczególną rolę – i to m.in. dzięki ich szerokiemu zastosowaniu w projekcie zrodził się pomysł na niniejszą pracę. Szczegóły tego zagadnienia zostaną przedstawione w sekcji 2.4.

2.3. Wiktionary – Wikisłownik

Jednym z największych projektów siostrzanych Wikipedii jest Wikisłownik, w wersji angielskiej (i wielu innych) noszący nazwę *Wiktionary* (`http://www.wiktionary.org`). Ten słownik internetowy nie rozwinął się jeszcze tak prężnie jak encyklopedia, zwłaszcza jeśli chodzi o polską edycję. Jest dziś jednak jednym z największych słowników w sieci, a w pewnych zastosowaniach stanowi najlepszy wybór. Dużą zaletą Wikisłownika jest jego wielojęzyczność – w tym samym serwisie znaleźć można hasła w ponad 250 językach. W przypadku niektórych z nich jest to praktycznie jedyny słownik internetowy lub nawet jedyny dostępny słownik w ogóle. Przykładem może być polski Wikisłownik, który zawiera prawdopodobnie jedyny polski słownik języka hawajskiego czy największe słowniki języków suahili i jidysz.

Wspomniane zostało zastosowanie szablonów do standaryzacji kodu źródłowego i struktury haseł. Trzeba jednak zaznaczyć, że dotyczy to wyłącznie haseł w obrębie jednej wersji ję-



Ilustracja 2.2: Polska edycja Wikisłownika

zykowej Wikisłownika, są one bowiem niezależne od siebie i od Fundacji. Wspólne dla wszystkich edycji jest jedynie oprogramowanie MediaWiki i umiejscowienie na serwerach Fundacji pod adresem `xxx.wiktionary.org`, gdzie zamiast `xxx` wstawiany jest dwu- lub trzyliterowy skrót języka (np. `pl.wiktionary.org` = język polski, `de.wiktionary.org` = język niemiecki, `sq.wiktionary.org` = język albański, `csb.wiktionary.org` = język kaszubski). Wszystkie kwestie organizacyjne w obrębie wersji językowej ustalane są w ramach dyskusji i głosowań przez internetową społeczność. Głosowania służą także wyborowi administratorów projektu, czyli użytkowników mających dodatkowe uprawnienia, spośród których najważniejsze to usuwanie i zabezpieczanie haseł oraz blokowanie użytkowników działających na szkodę projektu. W lipcu 2011 roku w angielskim Wikisłowniku działało aktywnie 76 administratorów, zaś w polskim – 14. Dla porównania Wikipedia w języku angielskim ma 1541 administratorów (niekoniecznie aktywnych), wersja polska natomiast 163.

Podobnie jak w przypadku Wikipedii, hasła w poszczególnych wersjach językowych Wikisłownika są łączone poprzez odnośniki interwiki, znajdujące się na dole lewego menu w większości haseł. W stosunku do encyklopedii występuje znacząca różnica w sposobie funkcjonowania tych linków. Wikipedia poprzez interwiki łączy artykuły na ten sam temat, często różniące się tytułem (polski artykuł *Kot domowy* odsyła do angielskiego *Cat* czy niemieckiego *Hauskatze*). W Wikisłowniku mechanizm interwiki łączy zaś strony o tym samym tytule, niezależnie od ich zawartości. Polskie hasło *kot* zawiera więc łącza do haseł o tytule *kot* w innych językach, które mogą, ale nie muszą zawierać m.in. objaśnienia polskiego znacze-

nia tego słowa. Z tego względu łatwo jest odnaleźć brakujące informacje o wybranym słowie w danym języku, jeśli skorzysta się z kilku edycji językowych Wikisłownika. Natomiast tłumaczenia tytułu hasła na inne języki znajdują się bezpośrednio w treści artykułu, w miejscu przeznaczonym dla takich informacji.

Choć poszczególne wersje Wikisłownika różnią się między sobą, wspólna jest najbardziej ogólna struktura hasła. Każde hasło jest podzielone na sekcje nagłówkami 2. stopnia (patrz tabela 2.2.1), podobnie jak w innych projektach Fundacji. W przypadku słownika sposób podziału artykułu jest jasno sprecyzowany i wspólny dla wszystkich edycji. W każdej z sekcji objaśnione jest znaczenie tytułowego hasła w innym języku. Przykładowo hasło *nie* w większości wersji językowych ma sekcje objaśniające znaczenie m.in. w języku polskim oraz języku niemieckim (*nigdy*). Właśnie ta ogólna cecha struktury haseł pozwala na częściową automatyzację niektórych często wykonywanych w trakcie edycji czynności.

Aplikacja opisana w niniejszej pracy przeznaczona jest dla polskiej edycji Wikisłownika. Konieczne jest zatem scharakteryzowanie specyfiki tego projektu (patrz sekcja 2.4). Ze względu na odmienność poszczególnych wersji językowych prawdopodobnie w innych nie będzie można wykorzystać wykorzystać aplikacji. Z pewnością może ona być dla nich jednak bardzo przydatna – duża część funkcji przez nią realizowanych może być używana niezależnie od specyfiki danej wersji, a budowa aplikacji pozwoli na ich wyodrębnienie.

2.4. Polska edycja Wikisłownika

2.4.1. Struktura hasła

Rozdział 3

Aspekty społecznościowe

3.1. Koncepcja *wiki*

3.2. Społeczność polskiej edycji Wikisłownika

3.3. Specyfika tworzenia aplikacji dla wikispołeczności

Rozdział 4

Opis implementacji

4.1. Wprowadzenie

4.2. Formularz edycyjny

4.3. Automatyzacja edycji hasła

4.4. Wdrożenie

Rozdział 5

Podsumowanie