POLSKIE BIBLIOTEKI CYFROWE 2008 • **str. 73-79**

Systemy organizacji wiedzy i ich rola w integracji zasobów europejskich bibliotek cyfrowych

ADAM DUDCZAK

Poznańskie Centrum Superkomputerowo-Sieciowe, Poznań
maneo@man.poznan.pl

Streszczenie

Od 2007 roku trwają prace mające na celu utworzenie Europeany – paneuropejskiego cyfrowego muzeum, archiwum i biblioteki. Głównym zadaniem Europeany jest udostępnienie dorobku europejskiej kultury i nauki w postaci cyfrowej. Prace te są realizowane w ramach wielu projektów europejskich takich jak EDLnet czy EuropeanaLocal. Europa jest kontynentem wielu narodowości, kultur i języków – stworzenie uniwersalnej, cyfrowej usługi dostępowej do tak różnorodnych zbiorów dziedzictwa kulturowego jest zadaniem trudnym. W niniejszym artykule omówiono wymagania funkcjonalne i założenia będące podstawą działania systemu pozwalającego na wyszukiwanie informacji zapisanej w wielu, często bardzo odmiennych, językach. Jednym z narzędzi pozwalających osiągnąć takie możliwości są systemy organizacji wiedzy (ang. *Knowledge Organisation System*). W artykule omówione zostaną podstawowe zagadnienia związane z budową tego typu struktur i planami ich wykorzystania zarówno na poziomie europejskim, w usługach takich jak Europeana, jaki i przy agregowaniu zasobów na poziomie poszczególnych krajów, gdzie również występują problemy związane z wielojęzycznością i różnorodnością zbiorów. Tu przykładem może być Federacja Bibliotek Cyfrowych.

Słowa kluczowe: systemy organizacji wiedzy, wielojęzyczność, agregacja zasobów, Europeana, Federacja Bibliotek Cyfrowych

Wstęp

Tysiąclecia historii Europy to ogromna różnorodność i bogactwo dziedzictwa kulturowego, jakie pozostawili po sobie zarówno wielcy artyści, przywódcy, jak i zwykli ludzie. Budując nowoczesne społeczeństwo oparte na informacji musimy być świadomi tego, jak dużym wyzwaniem jest ta różnorodność zasobów. Digitalizacja jest zaledwie pierwszym krokiem, który trzeba podjąć, aby „uwolnić" informację i udostępnić zabytki kultury europejskiej szerokiemu gronu odbiorców. Jednak cyfryzacja to nie wszystko, setki płyt DVD z zapisanymi cyfrowymi wtórnikami, zamknięte w szafach, w żadnej mierze nie przysłużą się poszerzaniu wiedzy na temat naszej historii, czy rozwojowi kultury i nauki. Konieczne jest stworzenie odpowiedniej infrastruktury pozwalającej na budowę cyfrowych repozytoriów, w ramach których udostępniane byłyby zdigitalizowane obiekty. Jest to zadanie niełatwe, ale wiele wskazuje, że wysiłki jakie podjęto w tym zakresie w ostatnich latach przyniosły oczekiwany skutek. Doskonałym przykładem może być tutaj unikalna w skali europejskiej sieć polskich bibliotek cyfrowych.

Na zasoby zgromadzone w polskich bibliotekach cyfrowych musimy spojrzeć w szerszym kontekście, w kontekście możliwości jakie stwarza ogólnoświatowa sieć informacyjna. Dokumentami w nich udostępnianymi są zainteresowani nie tylko polscy użytkownicy, ale również osoby, które nie znają języka polskiego. Osoby takie mogą poszukiwać informacji o swoich przodkach, które

znajdą w przedwojennych polskich gazetach bądź księgach adresowych. Mogą być zainteresowane cyfrową wersją unikatowego manuskryptu napisanego w ich ojczystym języku, który w wyniku historycznych perturbacji trafił do Polski i tu został zdigitalizowany. Turyści, którzy zwiedzają Polskę, mogą szukać informacji o jej zabytkach architektonicznych. Realizując te bardzo podstawowe zadania użytkownik może napotkać wiele różnego rodzaju przeszkód. Zagadnienia te zostały zidentyfikowane i opisane w ramach działań wielu projektów europejskich między innymi MultiMatch [1], MACS [2], CACAO [3] czy projektu EDL [4]. Efektywne rozwiązanie tych problemów jest szczególnie istotne w kontekście trwających od 2007 roku prac nad Europeaną, paneuropejskim cyfrowym muzeum, archiwum i biblioteką [5].

Celem niniejszego artykułu jest opisanie najważniejszych problemów związanych z udostępnianiem i wyszukiwaniem informacji zapisanej w wielu językach. Omówione zostaną podstawowe wymagania funkcyjne dla usługi cyfrowego repozytorium, dającego równorzędny dostęp do treści wielojęzycznych. Wytyczne te w większości zostały zaproponowane w ramach prac projektu EDLnet.

Przedstawione zostaną również pewne ilościowe dane świadczące, iż problem wielojęzyczności dotyczy również zasobów zgromadzonych w polskich bibliotekach cyfrowych. Dotychczasowe badania w zakresie udostępniania zasobów wielojęzycznych wskazują, że do rozwiązania niektórych problemów mogą posłużyć używane w bibliotekarstwie od wielu lat systemy organizacji wiedzy (SOW). W dalszej części artykułu omówione zostaną podstawowe definicję i pojęcia związane z SOW, opisane również będą wyniki prac Konsorcjum WWW (W3C), których celem jest wypracowanie spójnego standardu umożliwiającego zapisywanie struktur SOW w sposób czytelny dla programów komputerowych. Następnie omówione zostaną potencjalne obszary wykorzystania systemów organizacji wiedzy do rozwiązywania problemów związanych z różnymi aspektami udostępniania wielojęzycznych zasobów cyfrowych.

Podstawowe wymagania i problemy związane z udostępnianiem treści wielojęzycznych

W dokumencie [6] wymienione zostały najważniejsze wymagania funkcyjne, jakie będzie musiał realizować interfejs europejskiej biblioteki cyfrowej. Wymagania te zostaną zastosowane również przy budowaniu Europeany. Wspomniane wytyczne dotyczą zarówno samej strony WWW (tłumaczenie podstawowych mechanizmów nawigacyjnych), jak i zagadnień związanych z konstruowaniem indeksów wyszukiwawczych i realizowaniem procesu wyszukiwania w zgromadzonych zasobach.

Omawiane zagadnienia dotyczą w równej mierze repozytoriów cyfrowych, jak i ogólnych wyszukiwarek internetowych. Interfejs wyszukiwawczy Europeany i EDL pozwala użytkownikowi wyspecyfikować język, w jakim zadaje on zapytanie oraz, w przypadku polisemii, doprecyzować, czego dotyczy wydane zapytanie.

Zanim użytkownicy będą mogli przeszukiwać zgromadzone zasoby, konieczne jest stworzenie indeksów wyszukiwawczych. Jest to struktura, która pod wieloma względami przypomina tradycyjne biblioteczne katalogi kartkowe, czy znajdujący się na końcu niektórych książek skorowidz. W tradycyjnym skorowidzu dobór pojęć zależy od twórcy, człowieka, który dzięki posiadanej wiedzy i doświadczeniu jest w stanie wybrać z tekstu najważniejsze pojęcia. W przypadku dokumentów elektronicznych ręczne indeksowanie stosuje się bardzo rzadko, w większości zastosowań indeks wyszuki-

wawczy tworzy program komputerowy. Aby zbudować wydajny indeks wyszukiwawczy konieczne jest zgromadzenie specyficznej wiedzy na temat języka, w którym obiekt został stworzony oraz innych zasobów. Ponieważ reguły indeksowania mogą być wzajemnie sprzeczne, w przypadku różnych języków usługa wyszukiwawcza operująca na wielojęzycznych zasobach powinna tworzyć osobne indeksy dla każdego z obsługiwanych języków.

Przetwarzanie dokumentu w procesie tworzenia indeksu powinno uwzględniać morfologiczne cechy poszczególnych wyrazów, z których składa się dokument. Wiedza ta jest niezmiernie istotna w przypadku języków o bogatej fleksji, jakim jest na przykład język polski. Zignorowanie tej kwestii może spowodować, że wyszukanie dokumentów zawierających jakieś słowo okaże się niemożliwe.

Innym ważnym problemem jest dostępność list wyrazów pospolitych (ang. *stop words list*). Mianem wyrazów pospolitych określa się słowa, które nie posiadają samodzielnego znaczenia, ale występujące w tekście bardzo często np. „na", „to". Wyrazy pospolite są pomijane w procesie indeksowania, pozwala to znacznie zmniejszyć rozmiar tworzonego indeksu bez utraty jakości otrzymywanych wyników. W przypadku wielojęzyczności należy pamiętać, że słowo, które w jednym języku jest wyrazem pospolitym, w innym może być np. rzeczownikiem. Za przykład niech posłuży wyraz „but", który w języku polskim jest rzeczownikiem i nie znajduje się na liście wyrazów pospolitych, natomiast w języku angielskim jest przeważnie pomijany, jako mało znaczący.

Stworzenie zaawansowanych narzędzi umożliwiających automatyczne tłumaczenie dokumentów metodami tłumaczenia maszynowego (ang. *machine translation*, MT) lub tworzenia ich streszczeń (ang. *automatic summarization*) wymaga dostarczenia dodatkowych zasobów, pozwalających na przeprowadzenie analizy struktury syntaktycznej i semantycznej dokumentów.

W specyfikacji [6] określono także trzy podstawowe typy wyszukiwań:

1) zapytanie w określonym języku zwraca w wyniku tylko obiekty w tym języku,
2) zapytanie w określonym języku zwraca pasujące obiekty niezależnie od języka, w którym zostały one sporządzone,
3) zapytanie w określonym przez użytkownika języku jest tłumaczone na wszystkie wspierane języki i zwraca obiekty we wszystkich (wspieranych) językach.

W przypadku gdy dysponujemy oddzielnymi indeksami dla każdego z wspieranych języków i omówionymi dotychczas zasobami językowymi, realizacja dwóch pierwszych przypadków nie powinna sprawić żadnych trudności. Aby zrealizować trzeci ze scenariuszy, potrzebne nam będą dodatkowe zasoby: słowniki odwzorowujące język zapytania na pozostałe wspierane języki lub zewnętrzną usługę, która zrealizuje ten cel przy użyciu narzędzi MT.

Zalecenia [6] sugerują użycie Unicode dla potrzeb kodowania znaków, pomoże to w rozwiązaniu problemów z obsługą znaków specyficznych, występujących w niektórych alfabetach.

Odrębną kwestią jest możliwość łatwego wydawania zapytań w dowolnym języku. Poza wsparciem uniwersalnego kodowania znaków konieczne jest również dostarczenie swoistej wirtualnej klawiatury, która pozwoli na użycie znaków specyficznych dla alfabetów narodowych bez konieczności instalowania dodatkowych komponentów w komputerze użytkownika. Rozwiązanie tego typu dostępne jest w formularzu wyszukiwawczym The European Digital Library. Alternatywnym rozwiązaniem stosowanym z powodzeniem np. przez system dLibra jest przeprowadzenie konwersji liter niestandardowych do ich odpowiedników w alfabecie łacińskim np. *ą* zostaje zamienione na *a*.

Przeszukując zasoby wielojęzycznego repozytorium, po wydaniu zapytania zwrócona zostanie lista wyników. Aby w pełni wykorzystać zasoby, które się na niej znajdują, przewidziano możliwość tłumaczenia również wyników wyszukiwania. Ponieważ tłumaczenia oryginalnych dokumentów na język, w którym wydano zapytanie, z różnych względów może się okazać trudne w realizacji (np. brak OCR), w [6] założono, że w wersji podstawowej przetłumaczone zostaną przynajmniej opisy bibliograficzne odnalezionych dokumentów.

Wielojęzyczność zasobów zgromadzonych w polskich bibliotekach cyfrowych

Począwszy od 23 października, aż do 22 listopada 2008 roku strony Wielkopolskiej Biblioteki Cyfrowej (WBC, http://www.wbc.poznan.pl) odwiedziło 77 612 użytkowników posługujących się ponad 50 różnymi językami. Szczegółowy udział procentowy poszczególnych języków przedstawiono w tabeli 1. Na podstawie analizy wartość atrybutu „Język zasobów" można stwierdzić, iż w zbiorach WBC (ponad 70 tysięcy obiektów cyfrowych) znajdują się obiekty w 21 różnych językach, począwszy od polskiego na galijskim skończywszy. Podobną analizę przeprowadzono (dane z 5. 12. 2008 r.) dla innych bibliotek cyfrowych:

- Śląskiej Biblioteki Cyfrowej (http://sbc.org.pl) – 15 różnych językach,
- Biblioteka Cyfrowa Politechniki Warszawskiej (http://bcpw.bg.pw.edu.pl) – 6 różnych języków,
- Biblioteka Cyfrowa Politechniki Łódzkiej (http://ebipol.p.lodz.pl) – 5 różnych języków,
- Małopolska Biblioteka Cyfrowa (http://mbc.malopolska.pl) – 13 różnych wartości atrybutu „Języku zasobu".

Widać więc, że problemy związane z dostępem do wielojęzycznych zasobów dotyczą również polskich bibliotek cyfrowych.

Tabela 1. Udział procentowy języków używanych przez gości WBC

Język użytkowników	Udział procentowy
Polski	88,4%
Angielski	5,54%
Niemiecki	2,98%
Rosyjski	1,45%
Francuski	0,4%
Czeski	0,17%
Ukraiński	0,13%
Pozostałe	0,93%

Systemy organizacji wiedzy

Systemy organizacji wiedzy (ang. *Knowledge Organisation Systems*) to ogólne pojęcie, którego używa się dla określenia narzędzi takich jak słownictwo kontrolowane, tezaurus, taksonomia, klasyfikacja czy ontologia [7]. Ich najważniejszym zadaniem jest prezentowanie ogólnej wiedzy w sposób ustrukturalizowany i użyteczny.

Najprostszym z omawianych SOW jest słownictwo kontrolowane. Jest to zbiór unikalnych słów, z których każde posiada jednoznaczną definicję [7]. Jeżeli dane słowo posiada wiele znaczeń, wybierane to jest najczęściej stosowane, natomiast pozostałe zostają uzupełnione o jednoznacznie określający ich sens kwalifikator (por. [8]). Przykładem słownictwa kontrolowanego jest zbiór haseł Wikipedii, Medical Subject Headings (MeSH) czy słownik Katalogów Automatycznych Bibliotek Akademickich (KABA) [7].

Tezaurus rozszerza strukturę słownictwa kontrolowanego, dodając do niej informacje o terminach powiązanych w różny sposób. Nie istnieją żadne ograniczenia co do charakteru przechowywanych powiązań, do podstawowego zbioru relacji prezentowanych w tezaurusach należą: synonimia, hipernimia, hiponimia czy kolokacja. WordNet, jeden z najbardziej znanych zasobów tego rodzaju, definiuje nawet kilkanaście różnych typów relacji [10]. W kontekście bibliotek należy wspomnieć o amerykańskim tezaurusie Library of Congress Subject Headings (LCSH) czy jego francuskim odpowiedniku RAMEAU.

Warto podkreślić, że każdy rekord zapisany w słowniku KABA powinien zawierać nie tylko unikalne hasło i jego objaśnienie w języku naturalnym, ale również odwołanie do odpowiadającego hasła w LCSH i RAMEAU. Takie odwzorowanie może znacznie ułatwić poprawne automatyczne tłumaczenie opisów bibliograficznych.

Narzędzia te są doskonale znane od wielu lat, jednak dopiero rozwój technologii semantycznych (w tym prac nad siecią semantyczną) sprawił, że SOW przykuły uwagę takich instytucji jak konsorcjum WWW (W3C). Organizacja ta jest jednym z najważniejszych elementów w procesie tworzenia internetowych standardów. W roku 2002 w ramach działań konsorcjum rozpoczęto pracę nad opracowaniem standardu zapisu struktur takich jak tezaurusy czy słownictwa kontrolowane w postaci dokumentów XML. Wykorzystanie już istniejących standardów, takich jak RDF i RDF Schema ma sprawić, że SKOS (ang. *Simple Knowledge Organisation System*) będzie fundamentem sieci nowej generacji.

Obszary wykorzystania SOW w europejskich bibliotekach cyfrowych

Mimo ogromnego postępu metod tłumaczenia maszynowego wyniki automatycznej translacji wciąż pozostawiają wiele do życzenia. Dlatego wydaje się, iż użycie informacji i wzajemnych relacji, jakie istnieją między najbardziej znanymi różnojęzycznymi SOW, będzie nieodzowne w procesie integracji zbiorów europejskich. Mimo, iż sieć wzajemnych powiązań między słownictwami używanymi do opisu zasobów w różnych krajach jest bardzo rozbudowana, to tworzenie tego typu odwzorowań może wcale nie być sprawą trywialną. Można wyróżnić trzy podstawowe źródła trudności, jakie mogą wystąpić podczas dopasowywania pojęć pochodzących z dwóch różnych słownictw kontrolowanych: 1) niezgodności znaczeniowe, 2) różne poziomy szczegółowości, 3) brak pojęć odpowiadających. Zestawiając ze sobą dwa wyrazy musimy zwrócić uwagę na to, by ich znaczenie było jak najbardziej zbliżone. Nie zawsze da się osiągnąć stu procentową zgodność, przykładem może być niemiecki wyraz „Gemüse", którego polski odpowiednik „warzywa" ma szerszy zakres.

Sytuacje w jakich tworzymy odwzorowanie między dwoma słownictwami jest problemem bardziej ogólnym i dotyczy również różnych SOW w tym samym języku. Ze względu na dużą przydatność SOW, tworzone są często dedykowane struktury wiedzy, obejmujące swoim zakresem tylko pewną konkretną dziedzinę, przykładem tutaj może być MeSH. Są one przeważnie dość szczegó-

łowe i przyporządkowanie im terminów z struktur bardziej uniwersalnych może okazać się trudne, czy wręcz niemożliwe.

W zakresie tworzenia odwzorowań między słownictwami używanymi w różnych krajach podjęto już pracę w ramach projektu MACS (Multilingual Access to Subject). Jednym z celów tej inicjatywy było stworzenie odwzorowania między RAMEAU, LCSH i ich niemieckim odpowiednikiem SWD. Sukces tego projektu jest dobrym prognostykiem dla podobnych przedsięwzięć, które będą realizowane w przyszłości.

Zapewnienie zgodności na poziomie opisu cyfrowych obiektów pozwoli na bardziej zaawansowane wykorzystanie SOW. Twórcy systemów wyszukiwawczych starają się uzyskać od użytkownika jak najwięcej informacji dotyczących jego potrzeb informacyjnych. Wyszukiwanie nabiera cech konwersacji, wszystko po to, aby zwiększyć prawdopodobieństwo tego, że zwrócone wyniki pozwolą użytkownikowi zrealizować jego cele. Zastosowanie ustrukturalizowanej wiedzy może być również potencjalnie użyteczne dla systemów wyszukiwawczych [7]. W najprostszym przypadku wyszukiwanie to próba odnalezienia dokumentów, które zawierają wystąpienia pewnych słów wyspecyfikowanych przez użytkownika w formie zapytania. W większości przypadków użytkownicy wydają krótkie zapytania składające się z jednego, dwóch słów. System wyszukiwawczy dysponując wiedzą zgromadzoną w SOW mógłby podjąć próbę dopasowania zapytania wydanego przez użytkownika do któregoś z pojęć ze słownika. W przypadku gdy zapytanie byłoby terminem niejednoznacznym, wyszukiwarka musiałaby być w stanie określić, jaki jest właściwy sens zapytania wydanego przez użytkownika. W tej sytuacji możliwe są dwa rozwiązania:

- system zadaje użytkownikowi pytanie o to, które znaczenie problematycznego słowa miał na myśli,
- przeprowadzane jest kilka równoległych wyszukiwań dla każdego z możliwych znaczeń, na stronie z wynikami prezentowane są wyniki wszystkich wyszukiwań wraz z informacją, że wydane zapytanie było niejednoznaczne.

Zakładając, że system poprawnie rozpoznał, o który termin ze słownika pyta użytkownik, istnieje możliwość automatycznego przetłumaczenia wydanego zapytania oraz rozszerzenia go o terminy podrzędne (hiponimy). Rozważmy następujący przykład: użytkownik wydaje zapytanie „zwierzęta", system posiada informacje o tym, że słowa „ssaki", „ryby" i „ptaki" są pojęciami węższymi. Zapytanie jest rozszerzane o pojęcia podrzędne, a dzięki dostępności odwzorowań do słowników w innych językach może ono również zostać przetłumaczone. Opisy obiektów otrzymane w wyniku takiego wyszukiwania mogą zostać wyświetlone w języku, którym posługuje się użytkownik tak, aby mógł w pełni wykorzystać odnalezione informacje.

Wnioski końcowe

Problem wielojęzyczności to tylko jedna z przeszkód na drodze do zwiększenia interoperacyjności europejskich bibliotek cyfrowych. Skala przedsięwzięć, takich jak Europeana, rodzi również pytanie o przydatność SOW w sytuacji, gdy zgromadzona w nich wiedza będzie odzwierciedlać ogromną liczbę i różnorodność europejskich kolekcji dokumentów. To pytanie pozostanie bez odpowiedzi aż do chwili, gdy opracowane zostaną zalecenia dotyczące słownika wartości, jaki będzie wykorzystywała Europeana. Do tego czasu należy podjąć wysiłki, mające na celu ujednolicenie konwencji i słownictwa stosowanego do opisu obiektów cyfrowych w polskich bibliotekach cyfrowych.

Zgromadzenie odpowiedniej wiedzy, ustrukturalizowanie jej i zapisanie w postaci gotowej do automatycznego przetwarzania to spore wyzwanie.

Warto rozważyć wykorzystanie już istniejących narzędzi, kartoteka haseł wzorcowych KABA w 2002 roku zawierała 720 tysięcy rekordów, natomiast w grudniu 2007 roku już ponad 2,2 miliona. Z różnych względów KABA nie jest jednak używana przez wszystkie polskie biblioteki cyfrowe, również rozwiązania proponowane przez Bibliotekę Narodową nie są powszechnie akceptowane. Ogromne nadzieje należy wiązać z powołanym do życia w roku 2008 konsorcjum *Polskie Biblioteki Cyfrowe*, które ma dużą szansę stać się inicjatorem prac standaryzacyjnych w tym zakresie.

Preferowanym sposobem włączania zasobów do Europeany są krajowe agregatory metadanych. Ich rolą będzie nie tylko regularne zbieranie informacji o nowych obiektach, ale również dokonywanie przekształceń i czyszczenie udostępnianych dalej opisów bibliograficznych. W Polsce rolę takiego agregatora będzie pełnić serwis *Federacja Bibliotek Cyfrowych*.

Piśmiennictwo

[1] Multimatch, strona domowa projektu http://www.multimatch.org/, dostęp online 5. 12. 2008.
[2] Multilingual Access to Subjects (MACS), strona domowa projektu https://macs.vub.ac.be/pub/, dostęp online 5. 12. 2008.
[3] Cross-language Access to Catalogues and On-line Libraries (CACAO), strona domowa projektu http://www.cacaoproject.eu/, dostęp online 5. 12. 2008.
[4] European Digital Library Project (EDL), strona domowa projektu http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/edlproject/index.php, dostęp online 5. 12. 2008.
[5] Europeana, http://www.europeana.eu/, dostęp online 5. 12. 2008
[6] M. Dekkers, S. Gradmann, C. Meghini, EDLnet – D2.5 – Europeana Outline Functional Specification, wersja z 20. 08. 2008, dostęp online 5. 12. 2008.
[7] D. Daćko, *Zastosowanie ontologii do odkrywania wiedzy*, praca magisterska, Politechnika Poznańska, Poznań 2007.
[8] Wikipedia, Strona ujednoznaczniająca i zasady ujednoznaczniania, http://pl.wikipedia.org/wiki/Wikipedia:Strony_ujednoznaczniaj%C4%85ceEuropeana, dostęp online 5. 12. 2008.