POLSKIE BIBLIOTEKI CYFROWE 2008 • **str. 23-28**

CACAO – wielojęzykowy dostęp do katalogów on-line i bibliotek cyfrowych

JOLANTA MAZUREK

Biblioteka Kórnicka PAN
j.mazurek@bkpan.poznan.pl

Streszczenie

Projekt europejski CACAO (*Cross-language Access to Catalogues And On-line libraries*) jest 24-miesięcznym projektem celowym współfinansowanym w ramach programu eContentPlus. Partnerami projektu są biblioteki oraz instytucje komercyjne. Celem projektu jest umożliwienie wielojęzykowego dostępu do katalogów on-line i obiektów w bibliotekach cyfrowych. Tworzona infrastruktura wykorzystywać będzie techniki przetwarzania języka naturalnego w połączeniu z istniejącymi systemami pozyskiwania informacji w celu zintegrowania różnorodnych zasobów oraz rozwiązań technologicznych, będących w posiadaniu partnerów projektu. W artykule przedstawiono aktualny stan zaawansowania prac prowadzonych w projekcie w ramach 8 pakietów roboczych, z uwzględnieniem dotychczasowej realizacji zadań przez Bibliotekę Kórnicką PAN.

Słowa kluczowe: biblioteki cyfrowe, katalogi OPAC, projekt CACAO, program eContentplus, społeczeństwo informacyjne, wielojęzykowość Unii Europejskiej

Wstęp

Integracja oraz współpraca krajów w ramach Unii Europejskiej wniosła ze sobą zjawisko wielokulturowości, które ściśle powiązane jest z problemem wielojęzyczności. Polityka wspólnoty w tym zakresie jest jednoznaczna: Unia Europejska wspiera rozwój różnorodności językowej, promuje wielojęzykowość oraz partycypuje w ochronie ginących języków mniejszości etnicznych zamieszkujących Europę. W ostatnich latach rozwój technologii lingwistycznych przełamujących bariery językowe stał się priorytetem zadeklarowanym w ogłoszonej przez Unię Europejską *Inicjatywie i2010 na rzecz Europejskiego Społeczeństwa Informacyjnego* [1]. W tym duchu ogłoszono również konkurs eContentPlus [2] wspierający finansowo (całkowity wkład UEwynosi 149 000 000 euro) projekty realizowane w latach 2005-2008, których misją jest zwiększenie dostępu i ułatwienie dostępności do europejskich treści cyfrowych dla użytkowników pochodzących z różnych obszarów kulturowych. W 2007 roku w ramach tego konkursu unijna komisja przyznała finansowanie dla 17 projektów z różnych dziedzin życia, w tym dla 6 dotyczących bibliotek cyfrowych, a gromadzących zasoby z różnych obszarów nauki. Wśród nich znalazł się 24-miesięczny projekt celowy CACAO (*Cross-language Access to Catalogues And On-line libraries*) [3], którego realizację rozpoczęto w grudniu 2007 roku, a całkowity koszt projektu szacowany jest na 2 600 000 euro.

Opis projektu

Partnerzy

CACAO jest projektem interdyscyplinarnym, z pogranicza informatyki, lingwistyki i bibliotekoznawstwa, realizowanym przez 9 podmiotów reprezentujących te środowiska. Koordynatorem pro-

jektu jest europejski oddział firmy XEROX Research Centre Europe (Francja), który wraz z włoskimi firmami: CELI i GONETWORK oraz Instytutem Badań Lingwistycznych Węgierskiej Akademii Nauk (RIL) stanowią wparcie techniczne i technologiczne projektu. Dla CACAO instytucje te dostarczają narzędzi i metod z zakresu informatyki, lingwistyki, zagadnień ontologicznych oraz związanych z procesem przetwarzania języka naturalnego.

W realizacji projektu uczestniczy również 5 europejskich bibliotek, zdecydowanie heterogenicznych. Każda z nich jest biblioteką innego rodzaju, z charakterystycznymi zbiorami dla odpowiednich grup użytkowników. Pracują one w różnych systemach bibliotecznych, stosują różne formaty danych i różne klasyfikacje. Posiadają także różne typy zbiorów oraz dysponują własnymi doświadczeniami.

Biblioteka Uniwersytetu w Bolzano (FUB, Libera Università di Bolzano we Włoszech) posiada bogate doświadczenia w wielojęzyczności ze względu na fakt, iż studenci tego uniwersytetu są angielsko-, niemiecko- i włoskojęzyczni. W związku z tym np. proces katalogowania zbiorów odbywa się równolegle w 3 językach, co stanowi niezwykłą wartość dla projektu.

Podobnym doświadczeniem, związanym z wielojęzykowością, dysponuje jedna z największych bibliotek w Niemczech, Biblioteka Uniwersytecka w Getyndze (SUB, University of Goettingen, Goettingen State and University Library).

Francja reprezentowana jest przez Cité des Sciences et de l'Industrie (CSI) – rodzaj centrum kulturalnego skupiającego bibliotekę, muzeum i ośrodek kultury o profilu naukowo-technologicznym. Instytucja ta posiada wiele wydawnictw multimedialnych z tego zakresu, udostępnianych na potrzeby projektu CACAO.

W realizacji projektu uczestniczy również Węgierska Biblioteka Narodowa (HEL, National Széchényi Library). Od 1994 roku organizuje ona, poprzez Węgierską Bibliotekę Cyfrową (MEK) [4] dostęp do węgierskich treści edukacyjnych, naukowych i kulturowych. Obecnie MEK należy do najbardziej popularnych i znaczących repozytoriów cyfrowych w tym kraju.

Do projektu zaproszono również Bibliotekę Kórnicką PAN (BK PAN), znaną na świecie ze swoich cennych zbiorów stanowiących niewątpliwie ważny element europejskiego dziedzictwa narodowego. Nie bez znaczenia jest także fakt, iż od 2002 roku BK PAN czynnie uczestniczy w budowie Wielkopolskiej Biblioteki Cyfrowej [5]. Cyfrowe wersje zbiorów Biblioteki Kórnickiej stanowią blisko 40% zasobów udostępnianych przez polskie biblioteki cyfrowe poprzez portal Federacji Bibliotek Cyfrowych [6].

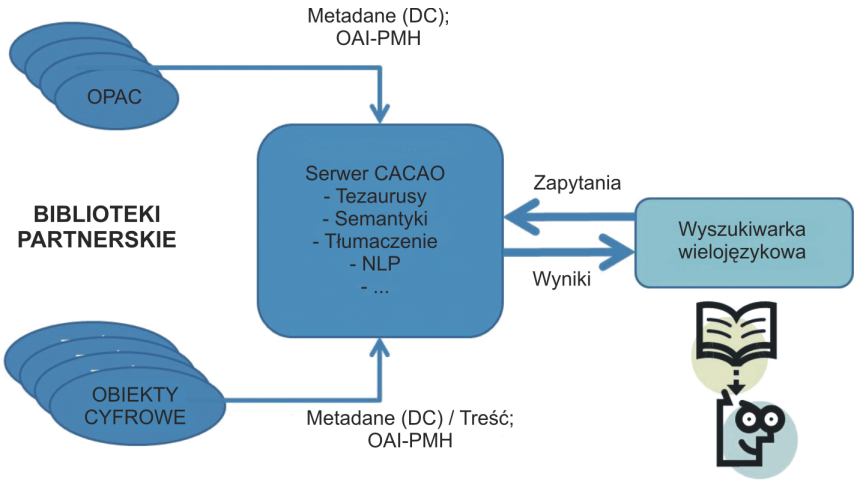
Skupienie tak zróżnicowanych bibliotek w ramach CACAO pozwala sądzić, iż bogate doświadczenia pochodzące z różnych środowisk bibliotekarskich, wsparte właściwymi działaniami partnerów technicznych i technologicznych, pozwolą na pomyślną realizację celów projektu. Wielojęzykowe katalogi on-line bibliotek oraz wielojęzykowe treści cyfrowe stanowią wspólną platformę, na której zorganizowano pracę w projekcie. Strona unijna dedykowana problemom wielojęzykowości wspólnoty [7] podaje, że 490 mln obywateli UE w 27 państwach członkowskich posługuje się 23 językami, uznanymi za języki oficjalne Unii. Wikipedia [8] podaje, że 47% ludności EU posługuje się językiem angielskim (jako ojczystym i obcym), a kolejno: 33% językiem niemieckim, 25% językiem francuskim, po 15% językiem włoskim i hiszpańskim, a 10% językiem polskim. Obszar działań projektu obejmuje 6 języków: angielski, niemiecki, włoski, francuski, węgierski i polski, które reprezentują 4 grupy językowe: germańską, romańską, ugrofińską i słowiańską.

Cel projektu CACAO

Głównym celem projektu CACAO jest utworzenie infrastruktury dla wielojęzykowego dostępu do katalogów on-line bibliotek i do publikacji udostępnianych w bibliotekach cyfrowych. W wyniku realizacji projektu powstanie serwer CACAO wyposażony w narzędzia, metody i oprogramowanie niezbędne do zrealizowania celów projektu. Zadaniem serwera jest przetwarzanie, rozszerzanie i wzbogacanie zapytań kierowanych przez użytkowników do systemu poprzez wykorzystanie włączonych na jego potrzeby narzędzi i technik z zakresu: ontologii, semantyki i zaawansowanych metod przetwarzania języka naturalnego. Na potrzeby projektu biblioteki partnerskie dostarczą poprzez protokół OAI-PMH [9] metadane w standardzie Dubline Core [10] ze swoich katalogów OPAC (ang. *Online Public Access Catalogue*) oraz z bibliotek cyfrowych, do serwera CACAO. Serwer wykorzysta te metadane do obsługi zapytań od użytkowników. Posiada on komponenty, które umożliwią realizację jego wielojęzykowych funkcji, a dodatkowo rozszerzą kontekst zapytania. W praktyce czytelnik zada pytanie serwerowi CACAO w jednym języku, a serwer posiadający wieloaspektowe funkcje oraz bazę danych o wielojęzykowych zasobach dokona tłumaczenia zapytania na inne języki i rozszerzenia jego kontekstu, wykorzystując zaadaptowane dla potrzeb Projektu wyspecjalizowane narzędzia. W wyniku takiego działania serwera użytkownik otrzyma odpowiedź nie tylko w swoim ojczystym języku, ale także w innych językach, odpowiedź będzie dodatkowo poszerzona (ryc. 1).

Wyniki projektu prezentowane będą na 3 portalach tematycznych: geograficznym, matematycznym i mediewistycznym. Poprzez te portale czytelnicy będą mogli dotrzeć do zagregowanych treści cyfrowych oraz do informacji zawartych w katalogach bibliotek.

Wypracowana w ramach projektu infrastruktura CACAO będzie w całości zaadaptowana przez The European Library (TEL) [11], która obecnie umożliwia dostęp do katalogów 48 narodowych bibliotek Europy (również Biblioteki Narodowej w Polsce) w 20 językach. Wykorzystanie tej infrastruktury w TEL znacznie wzbogaci funkcjonalność tej biblioteki, a czytelnikom ułatwi dostęp do opisów katalogowych z bibliotek nie tylko narodowych, ale również bibliotek innego typu, które posiadają opisy bibliograficzne w obcych, nieznanych czytelnikom językach.



Ryc. 1. Funkcje serwera CACAO

Zakres prac realizowanych w projekcie CACAO

Prace w CACAO prowadzone są w ramach 8 pakietów roboczych, w realizację których zaangażowani są wszyscy partnerzy projektu.

Pakiet 1 – Wielojęzykowy dostęp

Jest koordynowany przez firmę CELI i realizowany w pierwszych 12 miesiącach trwania projektu. Głównym celem prac prowadzonych w tym pakiecie jest zorganizowanie infrastruktury dla serwera CACAO. Ma ona zapewnić użytkownikowi końcowemu dostęp do informacji zgromadzonej w katalogach OPAC bibliotek oraz do treści cyfrowych we wszystkich dostępnych językach. Użyte metody i narzędzia muszą w sposób bezbłędny umożliwić czytelnikowi dostęp do szukanych przez niego informacji i wyeliminować wszystkie potencjalne dwuznaczności leksykalne. Realizacja głównych zadań tego pakietu dotyczyć będzie implementacji wszystkich narzędzi wzbogacających zapytania czytelników poprzez wykorzystanie dostępnych tezaurusów oraz metody, które w sposób jednoznaczny pozwolą na uzyskanie z systemu właściwych odpowiedzi w wielu językach. Szczegółowe prace obejmują m.in. opracowanie metod właściwej identyfikacji rzeczowników oraz jednoznacznego tłumaczenia systemów klasyfikacyjnych.

Pakiet 2 – Wielojęzykowe źródła

Jest koordynowany przez Xerox Research Centre Europe i realizowany w pierwszych 15 miesiącach projektu. Zadaniem tego pakietu jest dostarczenie wielojęzykowych źródeł zapewniających efektywne, wielojęzykowe wyszukiwanie w różnorodnych katalogach i zasobach cyfrowych, udostępnionych przez biblioteki współtworzące projekt. Głównym celem pakietu jest zaadaptowanie metod pozwalających na ciągłą rozbudowę wielojęzycznych słowników z wykorzystaniem zawartości kolekcji oraz wielojęzykowych zapytań wprowadzanych przez użytkowników.

Pakiet 3 – Infrastruktura projektu

Jest koordynowany przez Bibliotekę Uniwersytetu w Bolzano (FUB) i realizowany w pierwszych 12 miesiącach projektu. W realizację zadań tego pakietu zaangażowana jest także Biblioteka Kórnicka. Priorytetowym zadaniem pakietu jest wypracowanie odpowiednich standardów dla zorganizowania wspólnej infrastruktury dotyczącej niejednorodnych formatów oraz protokołów wymiany danych, stosowanych przez biblioteki biorące udział w projekcie. Utworzenie wspólnej ścieżki dostępu do różnych formatów oraz protokołów stanowi niezbędną podstawę do dalszych prac projektowych.

Biblioteka Kórnicka opracowuje swoje zbiory w formacie MARC 21 [12], a rekordy na potrzeby użytkowników Internetu udostępnia poprzez WebMAK. W przypadku rekordów katalogowych ujednoliconym formatem dla wszystkich partnerów projektu CACAO jest Dublin Core Simple. Dla potrzeb CACAO niezbędne okazało się zatem przemapowanie wszystkich rekordów z MARC21 do standardu Dublin Core Simple. Zatwierdzono także protokół dostępu do tych rekordów, jest nim OAI. W zakresie kolekcji obiektów cyfrowych i opisu metadanych formatem rekordów będzie Dublin Core Simple. W przeciwieństwie do rekordów katalogowych, Biblioteka Kórnicka PAN udostępnia tę informację w wymaganym formacie w Wielkopolskiej Bibliotece Cyfrowej. Tą drogą metadane trafiają także do CACAO.

W ramach pakietu planowana jest również organizacja współpracy z The Europan Library. Koordynator pakietu negocjuje obecnie z koordynatorem TEL warunki współpracy obu projektów, której efektem będzie zaadaptowanie przez TEL rozwiązań wypracowanych w projekcie CACAO.

Pakiet 4 – Dostęp dla użytkowników

Jest koordynowany przez XEROX Research Centre Europe. Prace rozpoczęły się w 6. miesiącu trwania projektu i potrwają do 21. miesiąca. Głównym celem działań jest organizacja dostępnego interfejsu dla użytkowników, umożliwiającego wielojęzykowe wykorzystanie zasobów dostarczonych przez biblioteki partnerskie CACAO. W ramach pakietu zrealizowane zostaną dwa interfejsy: prosty, który zapewni podstawową funkcjonalność wielojęzykowego przeszukiwania katalogów oraz interfejs zaawansowany rozbudowanej funkcjonalności.

Pakiet 5 – Agregacja zasobów

Jest koordynowany przez Bibliotekę Uniwersytecką w Getyndze (SUB) i realizowany od 7. do 24. miesiąca. W ramach tego pakietu dostarczone zostaną narzędzia umożliwiające łatwą integrację różnych bibliotek cyfrowych. Ponadto, jako pośredni wynik tego pakietu, utworzone zostaną wspomniane wyżej trzy portale dostępowe, w ramach których pojawią się zasoby dostarczone przez Bibliotekę Kórnicką. Aby zapewnić zgodność z wytycznymi dla CACAO, Biblioteka Kórnicka zorganizowała dostęp do tych metadanych poprzez protokół OAI-PMH do ponad 121 000 rekordów dla kolekcji różnego typu, tj: 42 500 starych druków, 29 900 nowych druków, 15 720 rękopisów, 500 czasopism do 1800 r., 7850 czasopism do 1800 r., 3250 gazet rękopiśmiennych, 500 dyplomów, 2580 dokumentów kartograficznych, 2500 materiałów z kolekcji tematycznej dotyczącej szachów oraz 15 900 mikrofilmów.

W ramach pakietu 5 Biblioteka Kórnicka (poprzez protokół OAI-PMH) udostępniła także metadane w standardzie Dublin Core Simple wybranych ok. 200 obiektów cyfrowych z Wielkopolskiej Biblioteki Cyfrowej. Tematycznie kolekcja ta zawiera dokumenty dotyczące literatury i historii średniowiecza i stanowi jeden z najstarszych i najcenniejszych zbiorów dziedzictwa kulturowego Europy, którego oryginały są przechowywane w Bibliotece Kórnickiej. Docelowo kolekcja ta ma stanowić produkt bazowy dla tworzonego w ramach CACAO Portalu Mediewistycznego. Poprzez ten portal czytelnicy będą mieli dostęp (w wielu językach) do kopii oryginałów m.in. takich materiałów jak: *Kronika Flandrii do roku 1384, Eneida* Wergiliusza, *Boska komedia* Dante Alighieri, rozpraw Seneki czy Cycerona itd. Dzięki dobrze przygotowanej infrastrukturze technicznej zastosowanej w Wielkopolskiej Biblioteki Cyfrowej, przygotowanie interfejsu OAI-PMH dla CACAO wymagało utworzenia jedynie dynamicznych zestawów OAI-PMH dla metadanych przydatnych z punktu widzenia projektu CACAO.

Pakiet 6 – Ocena i ewaluacja projektu

Jest koordynowany przez Bibliotekę Kórnicką PAN i realizowany w drugim roku trwania projektu. Celem pakietu jest oszacowanie trafności założeń i rozwiązań technicznych oraz metod zastosowanych w Projekcie, a także ocena satysfakcji użytkowników. W części dotyczącej rozwiązań technicznych zadaniem pakietu będzie opisanie możliwości zrealizowanego systemu pod kątem wykonywania zaawansowanych, wielojęzykowych rozszerzeń zapytań oraz jakości użytych słowników. Część prac związana z oceną zadowolenia użytkowników jest głównie poświęcona ewaluacji graficznego interfejsu użytkownika oraz jego integracji z podstawowymi warstwami systemu.

Pakiet 7 – Działania biznesowe

Jest koordynowany przez XEROX Research Centre Europe. Prace w ramach tego pakietu trwają przez cały okres realizacji i angażują wszystkich partnerów. Celem pakietu jest zaplanowanie

i zainicjowanie działań biznesowych skupionych na wynikach projektu. Głównym zadaniem jest wypracowanie dokumentu typu *business plan,* który pozwoli na podjęcie działań zapewniających trwałość wyników projektu po zakończeniu jego finansowania przez Unię Europejską.

Pakiet 8 – Promocja i rozpowszechnianie wyników

Jest koordynowany przez Cité des Sciences et de l'Industrie (CSI). Działanie prowadzone są przez cały czas trwania projektu i mają na celu organizację promocji Projektu oraz jego wyników (w fazie końcowej). W ramach działań przygotowane zostały materiały informacyjne na temat projektu (ulotki, strona WWW, prezentacja), a także artykuły publikowane na krajowych i międzynarodowych konferencjach i spotkaniach warsztatowych. Biblioteka Kórnicka zaprezentowała projekt w ramach konferencji „Polskie Biblioteki Cyfrowe" 25. 11. 2008 roku w Poznaniu. Projekt CACAO został przedstawiony na sesji poświęconej współpracy polskich instytucji w ramach projektów unijnych, obok projektów DRIVER, ENRICH i EuropeanaLOCAL.

Projekt CACAO zostanie zakończony z końcem 2009 roku. Powstałe wyniki, zgodnie z przyjętym planem biznesowym zostaną przedstawione użytkownikom. Czytelnicy uzyskają dostęp do europejskich treści cyfrowych w wielu językach. Polityka Unii Europejskiej ukierunkowana na zwiększenie dostępności różnokulturowej treści wciąż jest podtrzymywana i rozwijana. Nowe, zaproponowane przez Komisję Europejską możliwości współfinansowania wniosków w ramach kolejnego konkursu projektów w eContentplus, dotyczą kwoty 14 000 000 euro na rozwój narzędzi, standardów i metod ułatwiających dostęp do treści z obszaru różnych kultur. Różnorodność kulturowa w tego typu wnioskach, również spoza obszaru eContetnplus, stanowi niezwykle cenną wartość i ma istotny wpływ na decyzję o współfinansowaniu tego typu wniosków. Niezaprzeczalnym elementem współtworzącym dziedzictwo kulturowe Europy jest bogata i wartościowa kultura naszego kraju. Stąd duża szansa dla polskich instytucji na udział w projektach europejskich, a tym samym na wykorzystanie szansy promowania polskiej kultury i instytucji na arenie międzynarodowej.

Piśmiennictwo

[1] i2010 strategy, http://ec.europa.eu/information_society/eeurope/i2010/strategy/ (dostęp 8. 12. 2008 r.).
[2] The eContentplus Projects, http://ec.europa.eu/information_society/activities/econtentplus/projects/funded_projects/ (dostęp 8. 12. 2008 r.).
[3] The CACAO Project, http://www.cacaoproject.eu/ (dostęp 8. 12. 2008 r.).
[4] Hungarian Electronic Library, http://mek.oszk.hu/ (dostęp 8. 12. 2008 r.).
[5] Wielkopolska Biblioteka Cyfrowa, http://www.wbc.poznan.pl/ (dostęp 8. 12. 2008 r.).
[6] Federacja Bibliotek Cyfrowych, http://fbc.pionier.net.pl/ (dostęp 8. 12. 2008 r.).
[7] Languages of Europe, http://ec.europa.eu/education/languages/languages-of-europe/(dostęp 8. 12. 2008 r.).
[8] European Languages. *Wikipedia,* http://en.wikipedia.org/wiki/European_languages (dostęp 8. 12. 2008 r.).
[9] The Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/openarchivesprotocol.html (dostęp 8. 12. 2008 r.).
[10] Dublin Core Metadata Element Set, http://dublincore.org/documents/dces/ (dostęp 8. 12. 2008 r.).
[11] The European Library, http://theeuropeanlibrary.org/ (dostęp 8. 12. 2008 r.).
[12] MAchine-Readable Cataloging, http://www.loc.gov/marc/ (dostęp 8. 12. 2008 r.).