Atomowe usługi w środowisku rozproszonych bibliotek cyfrowych

Cezary Mazurek[1], **Tomasz Parkoła**[1], **Marcin Werla**[1]

Streszczenie: Niniejszy artykuł przedstawia opracowaną przez autorów propozycję podziału systemów bibliotek cyfrowych na zbiór atomowych usług sieciowych. Usługi takie pozwalają na uruchomienie bibliotek cyfrowych opartych na architekturze SOA. Przedstawione podejście zapewnia wysoką skalowalność usług ułatwiającą rozbudowę o nowe cechy i elastyczne dostosowanie usług do określonych wymagań funkcjonalnych. Umożliwia ono również wykorzystanie przez bibliotekę cyfrową innych systemów sieciowych, które są w stanie zapewnić świadczenie określonej, niezbędnej jej, usługi. W pracy wyróżniono cztery podstawowe usługi systemu biblioteki cyfrowej i opisano ich funkcjonalność. Następnie przedstawiono efekty implementacji zaproponowanej architektury w oprogramowaniu dLibra, które stanowi podstawę środowiska polskich bibliotek cyfrowych w sieci PIONIER.

Słowa kluczowe: biblioteki cyfrowe, rozproszone systemy przetwarzania danych, architektura zorientowana na usługi, usługi atomowe.

1. Wstęp

Biblioteki cyfrowe (BC) to systemy informatyczne tworzone w celu zbierania, długoterminowego przechowywania i udostępniania cyfrowych informacji. Jak pisze Michael Lesk, światowy specjalista w dziedzinie BC, „łączą one gromadzenie i organizowanie informacji, które od dawna prowadzone jest przez biblioteki i archiwa z cyfrową reprezentacją tych informacji możliwą dzięki komputerom" (Lesk, 1997). Systemy te, pierwotnie kojarzone przede wszystkim ze zbiorami zdigitalizowanych książek (Hart, 1992), mają dziś zdecydowanie szersze zastosowanie. Współczesne systemy BC pozwalają na przechowywanie obiektów cyfrowych w dowolnym formacie i umożliwiają opisanie tych obiektów przy pomocy zróżnicowanych schematów metadanych. Dzięki swojej elastyczności BC mogą być na przykład wykorzystywane do budowy repozytoriów naukowych zawierających zarówno publikacje naukowe opisujące rezultaty badań, jak i cyfrową postać materiałów źródłowych wykorzystanych w tych badaniach (np. serie pomiarów pochodzących z radioteleskopów, wirtualne modele zjawisk przyrodniczych czy skany i transkrypcje unikalnych zabytków piśmiennictwa). Inną płaszczyzną zastosowań BC są też systemy telemedyczne, gdzie BC wykorzystywane są jako repozytoria przypadków medycznych i baza wiedzy dla użytkowników tych systemów (Błaszczyński, 2006).

W Polsce prace badawczo-rozwojowe w zakresie BC prowadzone są co najmniej od drugiej połowy lat 90tych ubiegłego wieku (Mazurek, 1999). W ciągu ostatnich

[1] Poznańskie Centrum Superkomputerowo-Sieciowe, ul. Z. Noskowskiego 12/14, 61-704 Poznań.
Badania finansowane z grantu badawczego MNiSW nr 3 T11C 023 30.
{mazurek,tparkola,mwerla}@man.poznan.pl

pięciu lat w sieci PIONIER uruchomionych zostało około dwudziestu BC, które obecnie dają dostęp do ponad 70 000 obiektów cyfrowych. Tak dynamiczny rozwój systemów BC i przyrost gromadzonych w nich zasobów wymaga zastosowania specjalnych rozwiązań technologicznych, które umożliwią długotrwałe i wydajne działanie tych systemów oraz ich łatwą rozbudowę, skalowanie i integrację z innymi usługami sieciowymi.

W niniejszym artykule zawarto analizę systemów BC pod kątem możliwości ich implementacji w architekturze zorientowanej na usługi, propozycję usług BC dla takiej architektury oraz podsumowanie prototypowej implementacji takiej architektury w istniejącym środowisku rozproszonych bibliotek cyfrowych. Następny rozdział pracy zawiera opis tego środowiska utrzymywanego w sieci PIONIER, ze szczególnym uwzględnieniem oprogramowania wykorzystanego do jego budowy. Doświadczenia zebrane przy tworzeniu i utrzymywaniu platformy rozproszonych BC wraz z pracami opisanymi w (Dudczak, 2007) stanowią podstawę dla zaproponowanej w rozdziale 3 dekompozycji podstawowego zakresu funkcjonalnego biblioteki cyfrowej na cztery atomowe usługi sieciowe. Rozdział 4 zawiera wnioski z pierwszej próby dostosowania opisanego wcześniej oprogramowania do zaproponowanej koncepcji podziału BC na usługi. W ostatnim rozdziale pracy zawarto podsumowanie oraz proponowane kierunki dalszych prac.

2. Środowisko rozproszonych bibliotek cyfrowych w sieci PIONIER

2.1. Wprowadzenie

W październiku 2002 roku w Poznaniu, w ramach współpracy pomiędzy Poznańskim Centrum Superkomputerowo-Sieciowym i Poznańską Fundacją Bibliotek Naukowych uruchomiono Wielkopolską Bibliotekę Cyfrową - pierwszą tego typu usługę w polskim Internecie. WBC było też pierwszą BC opartą na rozwijanym przez PCSS od 1999 roku oprogramowaniu dLibra. W ciągu następnych lat w sieci Polski Internet Optyczny PIONIER uruchamiano kolejne BC. Dziś jest ich około 20 i liczba ta ciągle wzrasta. Wszystkie te biblioteki tworzą razem spójną platformę BC udostępniającą użytkownikom zaawansowane funkcje takie jak wyszukiwanie zasobów rozproszonych, unikalne identyfikatory czy wirtualne kolekcje obiektów cyfrowych. Poniżej krótko przedstawiono funkcjonalność systemu dLibra i jego architekturę oraz architekturę platformy rozproszonych BC.

2.2. Funkcjonalność systemu dLibra

Podstawową funkcją oprogramowania dLibra jest gromadzenie, organizacja, długoterminowe przechowywanie oraz prezentacja zasobów cyfrowych. Obiekty przechowywane w systemie dLibra mogą mieć dowolny format - zarówno PDF czy HTML jak i np. pliki audio/video. Po wprowadzeniu takich obiektów do BC mogą one być opisywane metadanymi, przy czym wyróżniamy tutaj metadane opisowe, administracyjne i techniczne. Struktura metadanych opisowych jest definiowana przez administratora na poziomie biblioteki cyfrowej w postaci hierarchicznego schematu

metadanych. Elementy znajdujące się na najwyższym poziomie domyślnie instalowanego schematu zgodne są ze standardem Dublin Core w wersji 1.1 (NISO, 2003). Metadane administracyjne opisują uprawnienia poszczególnych użytkowników do danego obiektu cyfrowego, jego lokalizację w strukturze biblioteki cyfrowej, przypisanie do kolekcji etc., natomiast metadane techniczne zawierają informacje na temat formatu i wewnętrznej struktury obiektu cyfrowego.

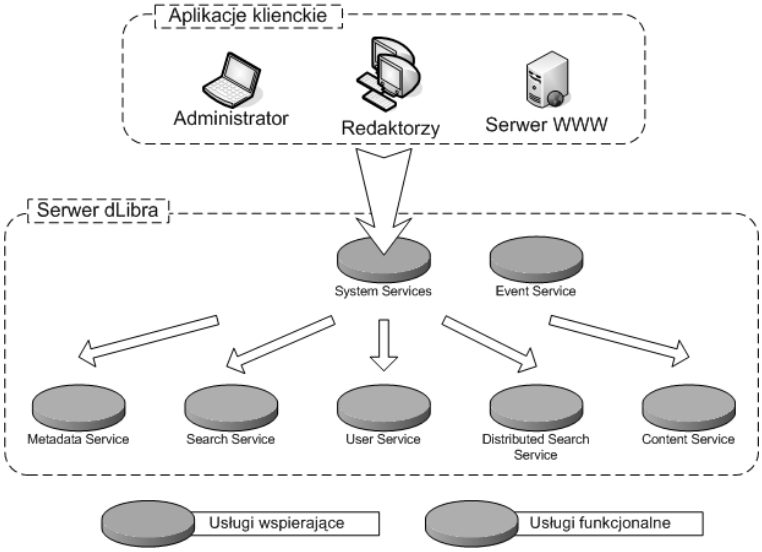
Podstawowym punktem dostępu do zasobów BC opartej na oprogramowaniu dLibra są strony WWW, dające uprawnionym użytkownikom m.in. dostęp do funkcji przeglądania zgromadzonych obiektów z uwzględnieniem podziału na kolekcje, przeszukiwania ich metadanych i treści (o ile da się ją przedstawić w postaci tekstowej) czy przeglądania indeksów wartości poszczególnych atrybutów. Oprogramowanie dLibra rozwijane jest również pod kątem integracji z innymi systemami informatycznymi. W tym zakresie dostępne jest m.in. pełne wsparcie dla protokołu OAI-PMH (Lagoze, 2004), wsparcie dla systemów jednokrotnego logowania, uwierzytelnianie i autoryzacja bazująca na protokole LDAP, kanały RSS oraz wymiana metadanych w formatach takich jak MARC, BibTeX czy RDF (lub też dowolny inny format oparty na języku XML). Dodatkowe informacje na temat zakresu funkcjonalnego systemu dLibra znaleźć można na oficjalnych stronach projektu: http://dlibra.psnc.pl/.

2.3. Architektura systemu dLibra

Oprogramowanie dLibra jest systemem wielowarstwowym o architekturze przedstawionej na rysunku 1. Podstawowym komponentem systemu jest serwer odpowiedzialny za realizację wszystkich funkcji dostępnych w systemie. Serwer ten złożony jest z siedmiu usług realizujących poszczególne funkcje systemu takie jak przechowywanie obiektów, przechowywanie metadanych, wyszukiwanie obiektów, zarządzanie użytkownikami czy zarządzanie komunikacją między usługami (Mazurek, 2005). Usługi serwera dLibra mogą być grupowane i uruchamiane na odrębnych komputerach (włącznie z możliwością uruchomienia każdej usługi na innym komputerze), co pozwala na skalowanie systemu w razie wzrostu liczby przechowywanych obiektów lub/i czytelników (Werla, 2006). W skład oprogramowania dLibra wchodzą również dwa komponenty stanowiące interfejs użytkownika. Komponenty te w celu realizacji swoich zadań wykorzystują funkcje poszczególnych usług serwera dLibra. Pierwszym z nich jest aplikacja WWW odpowiedzialna za prezentację zgromadzonych zasobów w Internecie. Drugim komponentem jest aplikacja okienkowa dzięki której możliwe jest wprowadzanie i zarządzanie obiektami cyfrowymi oraz administrowanie najważniejszymi elementami systemu (Parkoła, 2007).

2.4. Komunikacja w rozproszonym systemie bibliotek cyfrowych

BC uruchomione w sieci PIONIER tworzą razem rozproszony system stanowiący podstawę do tworzenia nowych zaawansowanych usług opartych o gromadzone w poszczególnych BC obiekty. W celu realizacji tych usług BC wymieniają między sobą zgromadzone metadane - każda z bibliotek udostępnia poprzez protokół OAI-PMH metadane lokalnie przechowywanych obiektów oraz przy pomocy tego samego



Rysunek 1. Wielowarstwowa architektura systemu dLibra

protokołu pobiera metadane zdalnych obiektów znajdujących się w innych bibliotekach cyfrowych. Wykorzystanie otwartego protokołu OAI-PMH pozwala dołączać do platformy BC oparte na zróżnicowanych rozwiązaniach technicznych - istotne jest tylko wsparcie dla tego protokołu. Wymiana metadanych między bibliotekami odbywa się okresowo, w sposób przyrostowy. W efekcie każda biblioteka cyfrowa posiada metadane wszystkich obiektów znajdujących się w całej sieci bibliotek. Takie podejście prowadzi do utrzymywania w poszczególnych BC platformy kopii metadanych obiektów z całej platformy, jednak dzięki temu możliwe jest realizowanie przeszukiwania i udostępnianie informacji o zdalnych obiektach bez konieczności odwoływania się do systemów, w których obiekty te są faktycznie przechowywane (Mazurek, 2006). Odwołanie to następuje dopiero w momencie próby pobrania treści tych obiektów. Obiekty cyfrowe identyfikowane są w platformie przy użyciu formatu identyfikatorów OAI. Format ten jest ograniczeniem specyfikacji URI i składa się z trzech części określających typ identyfikatora (stały - 'oai'), nazwę domenową konkretnej biblioteki cyfrowej (unikalną w skali całego Internetu) oraz lokalny identyfikator zasobu w ramach tej biblioteki. Takie podejście pozwala na łatwe i automatyczne tworzenie w poszczególnych BC identyfikatorów unikalnych w skali światowej.

3. Atomowe usługi bibliotek cyfrowych

Na przestrzeni ostatnich kilkudziesięciu lat wraz z rozwojem technologii zmieniała się również architektura systemów informatycznych. Kluczową rolę odegrały tutaj wzrost mocy obliczeniowej oferowanej przez komputery oraz rozwój technologii sieciowych umożliwiających rozproszenie systemów informatycznych na wiele

zdalnych lokalizacji. Pozwoliło to na ewolucję architektury systemów informatycznych od podejścia typu „mainframe", poprzez architekturę „klient-serwer", aż do powszechnej obecnie architektury wielowarstwowej. Systemy zrealizowane w tej architekturze charakteryzują się poziomym, „warstwowym" podziałem, w którym kolejne warstwy znajdują się coraz głębiej we wnętrzu systemu, tworząc tym samym swoisty stos, przez który przechodzą obsługiwane przez system żądania. Górną warstwą takiego stosu jest zazwyczaj interfejs użytkownika (obecnie najczęściej realizowany w technologii WWW), dolną warstwą jest system bazy danych, a pomiędzy nimi znajduje się jedna lub wiele warstw odpowiadających za realizację logiki biznesowej danego systemu.

Podejście wielowarstwowe w swojej czystej formie nie zapewnia jednak elastyczności i skalowalności oczekiwanej od współczesnych systemów informatycznych. Określenie konstrukcji poszczególnych „warstw" i sposobu ich komunikacji odbywa się zazwyczaj na etapie projektowania systemu i jest stałe i zamknięte w trakcie jego działania. Ponadto poszczególne warstwy zazwyczaj komunikują się tylko z warstwami bezpośrednio sąsiadującymi, przekazując dalej (i zazwyczaj przetwarzając) otrzymane informacje.

Obecnie coraz większą popularność zdobywa architektura zorientowana na usług (ang. *Service-Oriented Architecture, SOA*). W architekturze tej system informatyczny podzielony jest na część pełniącą rolę koordynatora usług oraz na usługi zapewniające funkcje niezbędne do realizacji logiki tego systemu. Koordynator usług to zazwyczaj część systemu będąca w stosunkowo bezpośrednim kontakcie z końcowym użytkownikiem (którym może być również inny system informatyczny), odpowiedzialna za obsługę żądań tego użytkownika przy pomocy dostępnych usług. Usługi te natomiast są odrębnymi rozproszonymi komponentami, które mogą być wykorzystywane przez wielu różnych koordynatorów (Lublinsky, 2007). Tak więc w takim podejściu unikalny charakter danego systemu informatycznego skojarzony jest z koordynatorem usług tego systemu, a dokładniej z rodzajami usług wykorzystywanych przez koordynatora i realizowanym przez niego sposobem interakcji czy przepływu danych między tymi usługami (ang. *choreography*, *orchestration*). Wybór poszczególnych usług przez koordynatora realizowany jest przy pomocy dodatkowego komponentu systemu - rejestru usług. Rejestr ten zawiera zazwyczaj informacje o dostępnych usługach wraz z ich charakterystykami (opcjonalnie). Dostawca usług zgłasza swoje usługi w rejestrze, a koordynator (konsument) usług kontaktuje się z rejestrem w celu pobrania informacji o usłudze, która jest w stanie zrealizować potrzebną w danej chwili funkcję.

Architektura SOA umożliwia zatem współdzielenie usług pomiędzy różnymi koordynatorami oraz dynamiczne dobieranie usług przez koordynatorów na potrzeby realizacji poszczególnych żądań. Kryteria doboru usług mogą mieć charakter:

- funkcjonalny - np. w portalu biologii obliczeniowej poszukiwana jest usługa *będąca w stanie dokonać asemblacji sekwencji DNA*;
- niefunkcjonalny - np. w systemie przechowywania nagrań z kamer monitoringu policyjnego poszukiwana jest usługa przechowywania plików, *która przesyła i przechowuje pliki w sposób zaszyfrowany z kluczem o długości co*

najmniej 1024 bitów i jest w stanie przechować 2 TB danych przez najbliższe 72 godziny.

Systemy BC ze względu na swój uniwersalny charakter (Kosiedowski, 2004) wydają się być doskonałymi kandydatami do implementacji w architekturze SOA. Możliwość wielokrotnego użycia pewnych usług składowych BC powinna pozwolić na zmniejszenie nakładu pracy przy tworzeniu nowych systemów tego typu oraz znacznie ułatwić utrzymanie i skalowanie istniejących obecnie dużych instalacji. Ustalenie ustandaryzowanych interfejsów takich usług oraz protokołów komunikacyjnych powinno również ułatwić wykorzystanie tych usług w innych systemach oraz wykorzystanie jako elementy składowe BC usług pierwotnie stworzonych dla innych systemów. W ramach prac opisanych w (Dudczak, 2007) dokonano analizy i porównania funkcjonalności najpopularniejszych na świecie pakietów oprogramowania do budowy BC, takich jak Fedora (http://fedora.info/), DSpace (http://www.dspace.org/) czy Open DLib (http://www.opendlib.com/). Analiza ta miała na celu wyszukanie wspólnych cech funkcjonalnych występujących w tych systemach i zestawienie ich z funkcjonalnością proponowaną przez dwa główne modele systemów BC: model OAIS (CCSDS, 2002) oraz model DELOS (Candela, 2006). W efekcie wyróżniono 4 podstawowe grupy funkcjonalne: przechowywanie obiektów cyfrowych, przechowywanie metadanych i adnotacji dotyczących tych obiektów, przeszukiwanie metadanych i adnotacji oraz tworzenie złożonych obiektów cyfrowych i powiązań między obiektami. Dalsza analiza wykorzystania tych funkcji doprowadziła do zaproponowania czterech następujących atomowych usług bibliotek cyfrowych:

1. Usługa przechowywania obiektów cyfrowych - powinna umożliwiać przechowywanie obiektów cyfrowych dowolnego typu i ich wersjonowanie. Jest to podstawowa i właściwie jedyna funkcja jaką ta usługa powinna spełniać. Dzięki takiemu założeniu możliwe będzie wykorzystanie w roli tej usługi wielu istniejących już usług sieciowych, takich jak serwery FTP/gridFTP, WebDAV czy nawet systemy wersjonowania takie jak CVS - oczywiście przy wykorzystaniu odpowiednich usług lub warstw pośredniczących odpowiedzialnych za dopasowanie interfejsów i translację specyficznych protokołów komunikacyjnych. Większość systemów BC wychodzi obecnie z założenia, że format przechowywanego materiał nie jest istotny z punktu widzenia usługi przechowującej. Staje się on ważny dopiero, gdy klient - inna usługa (np. wyszukiwawcza) lub użytkownik - próbuje z tego materiału korzystać. Jednak wtedy jest to kwestia odpowiedniego wsparcia dla danego formatu właśnie po stronie klienta. Problemu nie stanowią tutaj również dane udostępniane użytkownikom końcowym zazwyczaj poprzez strumieniowanie, gdyż można to zlecić dodatkowej usłudze, która będzie w stanie odczytać i strumieniować konkretne formaty obiektów cyfrowych (np. Real Audio) - jest to więc tylko kwestia dostępności i doboru odpowiednich usług przez koordynatora usług danego systemu.
2. Usługa przechowywania metadanych - powinna umożliwiać przechowywanie dowolnych metadanych powiązanych z obiektami cyfrowymi. W ogólności dla

każdego obiektu cyfrowego w systemie BC powinno byś możliwe przechowanie dowolnej liczby dowolnie zróżnicowanych zestawów metadanych. Stopień złożoności obsługiwanych formatów metadanych powinien być równy złożoności struktur jakie da się zapisać przy pomocy języka XML. Dodatkowo oparcie formatów metadanych w tej usłudze o język XML pozwoli na wykorzystanie do ich przetwarzania tak ogólnych mechanizmów jak XQuery i XPath oraz na wykorzystanie do implementacji takiej usługi coraz popularniejszych baz danych XML lub relacyjnych systemów baz danych wspierających przechowywanie danych w formacie XML (jak np. ostatnie wersje systemów baz danych firm Oracle i IBM oraz darmowe rozwiązania takie jak Xindice - http://xml.apache.org/xindice/). Warto tutaj również zwrócić uwagę na przechowywanie metadanych obiektów cyfrowych skojarzonych dodatkowo z konkretnymi użytkownikami BC, czyli adnotacji tych użytkowników dotyczących obiektów cyfrowych. W związku z rosnącą popularnością technologii Web 2.0, taką funkcjonalność udostępnia obecnie coraz więcej systemów BC.

3. Usługa kompozycji obiektów cyfrowych i tworzenia powiązań między nimi - powinna umożliwiać tworzenie relacji pomiędzy obiektami cyfrowymi (zarówno powiązań równorzędnych, jak i powiązań typu nadrzędny/podrzędny) oraz tworzenia grup powiązanych obiektów. Tak zdefiniowany mechanizm powinien zapewnić możliwość budowania złożonych obiektów cyfrowych oraz tworzenia kolekcji obiektów - obydwie te funkcje występują bardzo często w istniejących systemach BC. Przykładem mogą być tutaj zarówno kolekcje tematyczne (np. filmy związane z drugą wojną światową czy publikacje z dziedziny biologii obliczeniowej), jak i kolekcje obiektów cyfrowych danego formatu (kolekcja plików audio, kolekcja dokumentów HTML) czy wreszcie np. wyniki serii pomiarowych reprezentowane w BC jako złożone obiekty cyfrowe.

4. Usługa wyszukiwania obiektów cyfrowych - usługa ta powinna umożliwiać przeszukiwanie określonego zbioru obiektów cyfrowych danego typu lub zbioru metadanych danego formatu. Usługa ta na bazie zapytania powinna wygenerować listę referencji do obiektów cyfrowych, które spełniają to zapytanie. Usługa wyszukiwania obiektów cyfrowych jest niezbędna w systemach przechowujących duże ilości danych, jednak w przeciwieństwie do trzech poprzednich usług, jej implementacja będzie zapewne często charakterystyczna dla danej BC. O ile można np. stworzyć usługę przechowującą obiekty cyfrowe w dowolnym formacie, o tyle praktycznie niemożliwe jest stworzenie usługi umożliwiającej przeszukiwanie obiektów cyfrowych dowolnego typu - zasady budowania indeksów wyszukiwawczych na bazie obiektów cyfrowych czy ich metadanych oraz język zapytań będą tutaj musiały być przedmiotem ścisłych ustaleń przygotowywanych w pewnym z góry założonym kontekście związanym z konkretnym zastosowaniem systemu BC (np. repozytorium artykułów naukowych z funkcją przeszukiwania dokumentów tekstowych PDF oraz ich metadanych w schemacie Dublin Core, przy wykorzystaniu języka zapytań CQL).

Powyższe usługi powinny pozwolić na skomponowanie systemu BC dającego podstawową funkcjonalność jakiej oczekuje się od tego typu systemów. Przedstawione opisy usług skupiają się na ich cechach funkcjonalnych, pominięte zostały takie aspekty jak np. autoryzacja dostępu. Założono, że będzie ona realizowana w ramach wewnętrznych mechanizmów koordynatora usług lub też będzie realizowana w oparciu o zewnętrzny system/usługę i w związku z tym nie będzie miała wpływu na rozważane tutaj funkcje poszczególnych usług. Należy jednak pamiętać, że w kontekście konkretnej implementacji BC bardzo istotne staną się także aspekty niefunkcjonalne. Będą to zarówno kwestie związane np. ze wspomnianą autoryzacją czy specjalnym wsparciem dla wybranych formatów obiektów cyfrowych czy formatów metadanych, ale również aspekty takie jak długoterminowe przechowywanie obiektów cyfrowych (w tym np. automatyczna migracja między formatami/wersjami formatów danych), wydajność (może się np. okazać, że rozdzielenie wyszukiwania i przechowywania metadanych nie jest możliwe ze względów wydajnościowych konkretnej implementacji) czy bezpieczeństwo (brak możliwości wykorzystania zewnętrznych usług ze względu na konieczność ścisłej ochrony dostępu do poufnych danych gromadzonych w danej BC). Są to problemy, które trudno precyzyjnie zdefiniować i rozwiązać w oderwaniu od konkretnego zestawu wymagań i środowiska implementacyjnego. Następny rozdział pracy zawiera analizę możliwości zastosowania zaproponowanego powyżej podziału do modyfikacji architektury wykorzystywanego w wielu polskich BC środowiska dLibra.

4. Implementacja atomowych usług biblioteki cyfrowej w środowisku dLibra

Jak wspomniano wcześniej, system dLibra jest typowym systemem wielowarstwowym (patrz rysunek 1). Górną warstwę stanowią aplikacje klienckie, warstwa środkowa to serwer systemu odpowiadający za realizację głównych funkcji systemu, a dolną warstwę stanowi relacyjna baza danych. Sam serwer systemu dLibra podzielony jest na następujące usługi:

- Content Service - odpowiada za przechowywanie wersjonowanej treści obiektów cyfrowych,
- Metadata Service - odpowiada za przechowywanie metadanych i informacji o strukturze całej BC i poszczególnych obiektów oraz powiązań między obiektami,
- Search Service - odpowiada za indeksowanie i przeszukiwanie treści (tekstu) i metadanych lokalnych obiektów cyfrowych oraz metadanych zdalnych obiektów cyfrowych zgromadzonych przez usługę Distributed Search Service,
- User Service - odpowiada za zarządzanie użytkownikami i uprawnieniami do poszczególnych elementów BC,
- Distributed Search Service - odpowiada za pobieranie i przechowywanie informacji o obiektach cyfrowych znajdujących się w zdalnych BC.

Dodatkowo wyróżnione są jeszcze dwie usługi pomocnicze umożliwiające współpracę pozostałych usług:

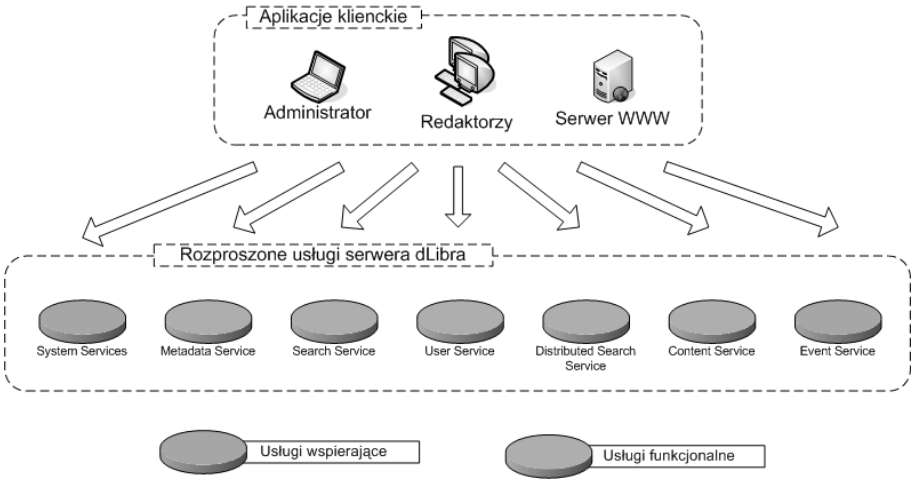
- Event Service - odpowiada za asynchroniczną komunikację usług i powiadamianie o zdarzeniach,
- System Service - jest punktem dostępowym do systemu i pośredniczy w wykorzystywaniu wszystkich innych usług serwera dLibra.

Zestawienie powyższych usług z funkcjonalnością czterech zaproponowanych usług atomowych przedstawia się następująco:

1. Usługa przechowywania obiektów cyfrowych - w przybliżeniu realizowana przez Content Service,
2. Usługa przechowywania metadanych - w przybliżeniu realizowana przez Metadata Service,
3. Usługa kompozycji obiektów cyfrowych i tworzenia powiązań między nimi - w przybliżeniu realizowana przez Metadata Service,
4. Usługa wyszukiwania obiektów cyfrowych - w przybliżeniu realizowana przez Search Service.

Jak widać usługi nr 2 i 3 w przypadku systemu dLibra realizowane są przez jedną usługę. Jest to w przypadku systemu dLibra uzasadnione specjalnymi wymaganiami związanymi z realizacją funkcjonalności dziedziczenia metadanych w strukturach obiektów cyfrowych. Funkcjonalność ta wymusza silne powiązanie pomiędzy wspomnianymi dwiema usługami. Nadmiarowo w stosunku do zaproponowanego modelu zdefiniowana jest w systemie dLibra usługa Distributed Search Service, z którą powiązana jest też część funkcjonalności usługi Search Service. Usługa Event Service jest również nadmiarowa, jednak realizuje ona kwestie niefunkcjonalne i jest związana ze sposobem implementacji mechanizmów komunikacji pomiędzy usługami w systemie dLibra. Usługa System Service pełni częściowo funkcję rejestru usług, jednak przez to, że pośredniczy ona w dostępie do funkcjonalności pozostałych usług, nie jest ona typowym rejestrem SOA. Fakt pośredniczenia powoduje, że serwer systemu dLibra zamiast dynamicznego zbioru usług jest z punktu widzenia aplikacji klienckich tylko kolejną warstwą tego systemu.

W ramach prototypowej implementacji zaproponowanego modelu usług w systemie dLibra zdecydowano się na wprowadzenie dwóch głównych modyfikacji. Pierwsza z nich polegała na zmianie roli usługi System Service. Zrezygnowano z pośredniczenia przez tą usługę we wszystkich wywołaniach funkcji serwera. Zamiast tego usługa ta stała się typowym rejestrem usług - gromadzi ona informacje o usługach danej BC i udostępnia je na żądanie. Dzięki temu aplikacje klienckie systemu dLibra stały się, w rozumieniu SOA, koordynatorami usług. Teraz w celu obsługi żądań użytkowników aplikacje klienckie łączą się bezpośrednio z wybranymi usługami systemu, pobrawszy uprzednio aktualne adresy tych usług z usługi System Service.



Rysunek 2. Architektura systemu dLibra po prototypowej implementacji modelu atomowych usług

Druga wprowadzona modyfikacja systemu dLibra polegała na przeniesieniu funkcjonalności przeszukiwania metadanych obiektów ze zdalnych BC z usługi Search Service do usługi Remote Search Service. Dzięki temu usługa Remote Search Service stała się usługą w pełni niezależną i atomową. Co więcej, ten krok wraz ze zmianą roli usługi System Service pozwolił na uczynienie usługi Remote Search Service usługą opcjonalną. Obecnie, gdy aplikacje klienckie stwierdzą, że brak w rejestrze usług aktywnej usługi Remote Search Service ukryją one elementy interfejsu z tą usługą związane. Dzięki wprowadzonym zmianom możliwe jest również współdzielenie jednej usługi Remote Search Service przez kilka BC opartych o system dLibra. Pozwala to na zmniejszenie stopnia duplikacji metadanych w ramach platformy bibliotek cyfrowych, o którym była mowa w punkcie 2.4.

Ostatecznym efektem wprowadzonych zmian była transformacja architektury systemu dLibra do postaci przedstawionej na rysunku 2.

5. Podsumowanie

W ramach prac przedstawionych w niniejszym artykule zaproponowano podział podstawowej funkcjonalności systemów BC na cztery atomowe usługi. Takie podejście umożliwia elastyczne konstruowanie i rozbudowywanie tych systemów oraz pozwala na wykorzystanie do ich budowy innych usług sieciowych, nie tworzonych pierwotnie z myślą o BC. Zaproponowany podział na usługi testowo zaimplementowano w oprogramowaniu dLibra wykorzystywanym do budowy około 20 bibliotek cyfrowych w Polsce. W efekcie udało się przetransformować architekturę tego oprogramowania do architektury typu SOA.

Kontynuacja opisanych prac polegać będzie przede wszystkim na analizie otwartych protokołów internetowych, które mogą zostać wykorzystane do komunikacji z

poszczególnymi usługami systemu BC. Wybór lub ewentualna definicja takich protokołów, a następnie ich implementacja umożliwi budowanie bibliotek cyfrowych w oparciu o usługi pochodzące od różnych dostawców, realizowane w różnych technologiach. W takiej sytuacji konieczne będzie określenie parametrów opisu usług BC, dzięki którym koordynatorzy usług będą w stanie dynamicznie decydować z jakiej instancji usługi danego typu chcą skorzystać. Takie możliwości wraz z dedykowanymi dla BC algorytmami odkrywania i kompozycji usług pozwolą na szybkie tworzenie nowych BC i zwiększą wykorzystanie funkcjonalności i zasobów bibliotek już istniejących.

Literatura

Błaszczyński, J., Kosiedowski, M., Mazurek, C. and Wilk, Sz. (2006) Ontologies for Knowledge Modeling and Creating User Interface in the Framework of Telemedical Portal. *In H. Stormer, A. Meier, M. Schumacher (eds): Proceedings of the European Conference on eHealth 2006 (ECEH'06). Fribourg, Switzerland, October 12-13, 2006.* Lecture Notes in Informatics P-91, 275-286. Gesellschaft für Informatik, Bonn.

Candela, L. and Castelli, D. (2006) *Reference Model for Digital Library Management Systems.* DELOS.

CCSDS (2002) *Reference Model for an Open Archival Information System (OAIS).* CCSDS.

Dudczak, A., Heliński, M., Mazurek, C., Parkoła, T. and Werla, M. (2007) Analiza funkcjonalności wybranych modeli i systemów zarządzania bibliotekami cyfrowymi. *Zeszyty Naukowe Wydziału ETI Politechniki Gdańskiej.* Politechnika Gdańska.

Hart, M. (1992) *The History and Philosophy of Project Gutenberg.* Project Gutenberg.

Kosiedowski, M., Mazurek, C. and Werla, M. (2004) Digital Library Grid Scenarios. *Knowledge-Based Media Analysis for Self-Adaptive and Agile Multi-Media, Proceedings of the European Workshop for the Integration of Knowledge, Semantics and Digital Media Technology, EWIMT 2004, November 25-26, 2004, London, UK*, 189 - 196. QMUL. ISBN 0-902-23810-8.

Lagoze, C. and Van de Sompel, H. (2004) *The Open Archives Initiative Protocol for Metadata Harvesting.* Open Archives Initiative. `http://www.openarchives.org/OAI/openarchivesprotocol.html`.

Lesk, M. (1997) *Practical digital libraries: Books, bytes and bucks.* Morgan Kaufmann.

Lublinsky, B. (2007) *Defining SOA as an architectural style.* IBM developerWorks.

http://www-128.ibm.com/developerworks/architecture/library/ar-soastyle/

Mazurek, C., Stroiński, M. and Szuber, S. (1999) Digital Library for Multimedia Content Management. *Proceedings of ERCIM 9th DELOS Workshop on Digital Libraries for Distance Learning.* ERCIM. ISBN 2-912335-08-6.

Mazurek, C. and Werla, M. (2005) Distributed Services Architecture in dLibra Digital Library Framework. *Proceedings of 8th International Workshop of the DELOS Network of Excellence on Digital Libraries on Future Digital Library Management Systems.* DELOS.

Mazurek, C., Stroiński, M., Werla, M. and Węglarz, J. (2006) Metadata harvesting in regional digital libraries in PIONIER Network. *Campus-Wide Information Systems*, Vol. 23, No. 4, pp 241 - 253. Emerald Group Publishing Limited. ISSN 1065-0741. ISBN 1-84663-184-X.

NISO (2003) *Information and documentation - The Dublin Core metadata element set, ISO Standard 15836-2003.* NISO.

Parkoła, T. (2007) *Podręcznik użytkownika środowiska dLibra (wersja 3.0).* PCSS. http://dlibra.psnc.pl/biblioteka/publication/2.

Werla, M. (2006) Architektura oraz możliwości skalowania systemu dLibra. *Baza wiedzy projektu dLibra.* PCSS. http://dlibra.psnc.pl/.

Atomic services in distributed digital libraries environment

This paper describes a decomposition of digital library systems into atomic network services. Those services allow building of a fully functional digital library in a way which ensures its flexible functioning, scaling and extending with new features. Such approach also gives a possibility to use other functionally similar network services in the process of digital libraries creation. In this paper we distinguish four atomic digital library services and describe their functionality. Next we describe first effects of prototype implementation of the proposed architecture into the dLibra software used to create and maintain the platform of digital libraries in the PIONIER network.