# Wdrażanie regionalnych bibliotek cyfrowych w sieci PIONIER w oparciu o środowisko dLibra

**Cezary Mazurek, Maciej Stroiński, Marcin Werla**

Poznańskie Centrum Superkomputerowo – Sieciowe

ul. Noskowskiego 12/14

61-704 Poznań

{mazurek,stroins,mwerla}@man.poznan.pl

**Streszczenie:** System dLibra to oprogramowanie do budowy bibliotek cyfrowych. System ten jest rozwijany w Poznańskim Centrum Superkomputerowo – Sieciowym od 1999 roku. Jest on wykorzystywany jako podstawa regionalnych i akademickich bibliotek cyfrowych w Polsce. W naszym referacie chcielibyśmy przedstawić bliżej możliwości oprogramowania dLibra oraz zwrócić uwagę na te jego cechy, dzięki którym umożliwia ono tworzenie bibliotek cyfrowych o zróżnicowanych wymaganiach funkcjonalnych, takich jak sposób autoryzacji czytelników czy dokładność opisu metadanymi.

**Summary:** dLibra is a software system for building digital libraries. It has been developed in Poznan Supercomputing and Networking Center since 1999. It is used as a software platform in regional and academic digital libraries in Poland. In our article we want to introduce the functionality of the dLibra framework and emphasize these dLibra features that make possible creation of dLibra-based digital libraries with diverse functional requirements, like user authorization mechanisms or precision of metadata descriptions.

## 1. Wstęp

dLibra jest to pierwszy polski system do budowy bibliotek cyfrowych, rozwijany w Poznańskim Centrum Superkomputerowo – Sieciowym od 1999 roku. Jest on wykorzystywany w Wielkopolskiej Bibliotece Cyfrowej [1] (od października 2002 roku) oraz w Bibliotece Cyfrowej Politechniki Wrocławskiej [2] (od listopada 2004 roku). Obecnie na ukończeniu są również prace mające na celu uruchomienie kolejnych bibliotek cyfrowych: Biblioteki Cyfrowej Uniwersytetu Zielonogórskiego, Podlaskiej Biblioteki Cyfrowej [3] oraz Kujawsko – Pomorskiej Biblioteki Cyfrowej [4]. Dodatkowo w wielu miastach takich jak Łódź, Lublin czy Warszawa uruchomione są testowe instalacje systemu dLibra.

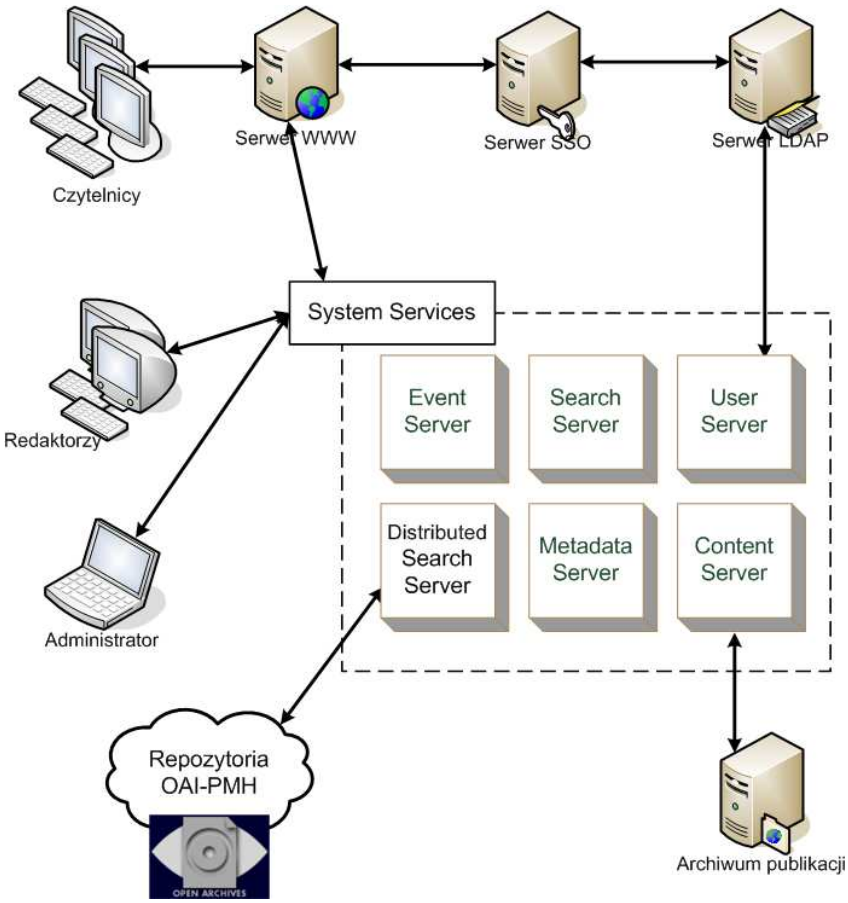
Taka liczba instalacji wymaga, aby oprogramowanie dLibra oferowało szeroki zakres funkcjonalności przydatnej użytkownikom i jednocześnie było elastyczne i wysoce konfigurowalne, w celu dostosowania się do potrzeb i wymagań konkretnej biblioteki cyfrowej. dLibra jest systemem rozproszonym i otwartym, który umożliwia czytelnikom dostęp do zawartości biblioteki poprzez interfejs WWW, dając równocześnie bibliotekarzom i administratorom zaawansowane narzędzia w postaci „Aplikacji Redaktora/Administratora". dLibra umożliwia przechowywanie i udostępnianie obiektów cyfrowych dowolnego typu – mogą to być zarówno dokumenty tekstowe w formatach takich jak HTML, PDF czy DjVu, jak i pliki audio czy video. Każdy z przechowywanych obiektów może być opisany metadanymi oraz przypisany do jednej lub wielu zdefiniowanych w danej bibliotece cyfrowej kolekcji. Użytkownicy systemu mają do dyspozycji zaawansowane mechanizmy tworzenia metadanych, takie jak słowniki wartości poszczególnych atrybutów czy obsługę formatów MARC i RDF. Czytelnicy mają możliwość przeglądania zawartości biblioteki oraz przeszukiwania metadanych poszczególnych obiektów oraz ich treści (dla określonych formatów).

Poniżej, w drugim rozdziale naszego artykułu prezentujemy architekturę systemu dLibra, dzięki której możliwe jest takie konfigurowanie dLibry, aby była ona w stanie zapewnić obsługę bibliotek cyfrowych o zróżnicowanych rozmiarach. Rozdział trzeci zawiera opis mechanizmów wykorzystywanych przy zarządzaniu metadanymi gromadzonymi w systemie. W rozdziale czwartym

przedstawiamy funkcje autoryzacji dostępu do treści cyfrowych zaimplementowane w systemie dLibra. Rozdział piąty to krótkie podsumowanie artykułu oraz przedstawienie planów dalszych prac.

## 2. Architektura systemu dLibra

Struktura systemu dLibra oparta jest na grupie rozproszonych współpracujących ze sobą usług (patrz Rys. 1). Usługi te dają razem pełną funkcjonalność systemu dLibra. Każda z usług może być uruchomiona na osobnym komputerze lub też może być jedną z usług tworzących grupę usług działających na jednym komputerze [5]. Każda z usług wymaga do swojego działania bazy danych. Usługi mogą współdzielić między sobą jedną bazę danych lub też wykorzystywać kilka niezależnych baz[1]. Dzięki temu uzyskujemy dużą skalowalność systemu – w przypadku wzrostu obciążenia biblioteki cyfrowej możliwe jest przeniesienie poszczególnych usług na osobne dedykowane serwery połączone siecią komputerową. Do wzajemnej komunikacji usługi systemu dLibra wykorzystują technologię Java RMI [6].




Rys. 1. Architektura systemu dLibra oparta o zestaw rozproszonych usług

W systemie dLibra wyróżniono następujące usługi:

- *Metadata Server* – daje możliwość definiowania, modyfikowania i usuwania atrybutów wykorzystywanych do opisu treści cyfrowej przy pomocy metadanych. Dodatkowo daje również dostęp do słowników i tezaurusów wartości poszczególnych atrybutów. Jest również odpowiedzialny za zarządzanie katalogami i kolekcjami biblioteki cyfrowej.

[1] System dLibra współpracuje z bazami danych Oracle, PostgreSQL oraz MySQL.

- *Content Server* – daje dostęp do treści gromadzonych w bibliotece cyfrowej. Treść przed przesłaniem do klienta może być kompresowana oraz szyfrowana. Usługa ta wykorzystywana jest również do przesyłania treści do biblioteki cyfrowej.
- *Search Server* – pozwala użytkownikom na przeszukiwanie zebranej treści i metadanych. Jest również odpowiedzialny za tworzenie indeksów wykorzystywanych podczas wyszukiwania.
- *Distributed Search Server* – jest wykorzystywany do pozyskiwania metadanych ze zdalnych instalacji systemu dLibra przy wykorzystaniu protokołu OAI-PMH. Serwer ten daje również użytkownikom możliwość przeszukiwania pozyskanych metadanych. Usługa ta może być wykorzystywana do przeszukiwania metadanych pobranych z każdego repozytorium udostępnionego przy pomocy protokołu OAI-PMH.
- *User Server* – zawiera wszystkie informacje związane z użytkownikami systemu i pozwala na autoryzację użytkowników. Pozwala on również na tworzenie grup użytkowników oraz na przydzielanie użytkownikom i grupom różnych uprawnień, od praw administracyjnych, do prawa przeglądania publikacji.

Komunikacja i współpraca pomiędzy powyższymi usługami odbywa się przy pomocy dwóch dodatkowych usług systemu dLibra. Pierwsza z nich to *System Services*. Usługa ta może być traktowana jako rejestr usług w pojedynczej instancji biblioteki cyfrowej. Umożliwia ona synchroniczną komunikację między usługami, jest odpowiedzialna za określanie adresów poszczególnych usług, łączenie z nimi i wzajemną autoryzację.

Druga z usług systemowych to *Event Server*. Usługa ta umożliwia innym usługom komunikację przy pomocy systemu asynchronicznego mechanizmu powiadomień o zdarzeniach.

## 3. Mechanizmy zarządzania metadanymi w systemie dLibra

Jak wspomniano wcześniej, za zarządzanie metadanymi odpowiedzialna jest usługa *Metadata Server*. Usługa ta umożliwia zdefiniowanie schematu atrybutów dostępnego w danej bibliotece cyfrowej. Schemat ten składać się może z dowolnej liczby atrybutów opisujących zasób cyfrowy, taki jak autor, nazwa, opis czy format zasobu. Predefiniowany w systemie dLibra zestaw atrybutów zgodny jest ze standardem Dublin Core Metadata Element Set (DCMES) [7] w wersji 1.1 i zawiera poniższe elementy:

- Tytuł - nazwa zasobu,
- Autor - instytucja lub osoba odpowiedzialna za zawartość zasobu,
- Temat i słowa kluczowe - tematyka zawartości zasobu,
- Opis - opis zawartości zasobu,
- Wydawca - instytucja lub osoba odpowiedzialna za publikację zasobu,
- Współtwórca - instytucja lub osoba, która wniosła wkład do zawartości zasobu,
- Data wydania - data związana z konkretnym wydarzeniem cyklu życia zasobu,
- Typ zasobu - charakter lub rodzaj treści zasobu,
- Format - sposób fizycznej lub cyfrowej prezentacji zasobu,
- Identyfikator zasobu - jednoznaczny identyfikator zasobu w pewnym kontekście,
- Źródło - odniesienie do zasobu, z którego wywodzi się ten zasób,
- Język - język zawartości zasobu,
- Powiązania - odnośniki do powiązanych zasobów,
- Zakres - zakres zawartości zasobu,
- Prawa - informacje o prawach dotyczących zasobu.

Zestaw ten można dowolnie dostosowywać poprzez zmianę, usuwanie oraz dodawanie nowych atrybutów. W celu zapewnienia zgodności dowolnego zdefiniowanego w systemie dLibra schematu atrybutów ze schematem DCMES, stworzono mechanizm ról atrybutów. W systemie zdefiniowano role atrybutów odpowiadające wszystkim elementom DCMES. Każdy atrybut zdefiniowany w

systemie dLibra może mieć przypisaną jedną rolę, przy żadna z ról nie może być przypisana do dwóch atrybutów. Dzięki rolom możliwe jest określenie, który z atrybutów w konkretnej instancji systemu dLibra odpowiada na przykład tytułowi publikacji. Możliwości te wykorzystywane są w wielu miejscach w systemie dLibra – na przykład w aplikacji czytelnika, do wyświetlania listy tytułów ostatnio dodanych publikacji.

Przy opisywaniu zasobu cyfrowego w systemie dLibra możliwe jest wprowadzenie wielu wartości dla każdego ze zdefiniowanych atrybutów. Możliwe jest również sporządzanie osobnych opisów dla dowolnej liczby języków. Dla każdego z atrybutów dynamicznie tworzony jest słownik wartości tego atrybutu. Słownik ten zawiera mechanizm umożliwiający łączenie zbliżonych znaczeniowo wyrazów w grupy. Mechanizm ten jest wykorzystywany do poprawy wyników wyszukiwania i może służyć zarówno dla obsługi typowych wyrazów bliskoznacznych jak i na przykład dla wprowadzenia do systemu dLibra kilku różnych pisowni nazwiska jednego autora.

W celu umożliwienia wymiany metadanych z zewnętrznymi systemami w dLibrze opracowano możliwości importu oraz eksportu metadanych. Możliwy jest eksport danych do formatu RDF oraz import danych z formatu RDF oraz MARC. Wykorzystywane są również protokół OAI-PMH [8] oraz format RSS [9].

## 4. Sposoby kontroli dostępu do treści cyfrowych

Zróżnicowane zastosowanie bibliotek cyfrowych spowodowało, iż system dLibra posiada rozbudowane możliwości dotyczące kontroli dostępu do gromadzonych treści cyfrowych. Jak wspomniano wcześniej, usługa *User Server* pozwala na definiowanie użytkowników oraz łączenie ich w grupy, a także na przyznawanie użytkownikom zróżnicowanych praw wykorzystywanych przy autoryzacji dostępu do zasobów. Prawa, które mogą być przyznane użytkownikom można podzielić na dwie kategorie: prawa administracyjne oraz prawa dotyczące konkretnych obiektów w strukturze biblioteki cyfrowej.

Istnieje sześć uprawnień administracyjnych, które dotyczą całej biblioteki cyfrowej. Są to:

- Zarządzanie kontami - pozwala na tworzenie, usuwanie i dokonywanie zmian w kontach użytkowników,
- Zarządzanie grupami - pozwala na tworzenie, usuwanie i dokonywanie zmian w grupach użytkowników,
- Zarządzanie atrybutami - pozwala na tworzenie, usuwanie i dokonywanie zmian w atrybutach,
- Zarządzanie wartościami atrybutów - pozwala na zarządzanie słownikiem synonimów,
- Zarządzanie kolekcjami - pozwala na tworzenie, usuwanie i dokonywanie zmian w kolekcjach,
- Zarządzanie aplikacją WWW – pozwala na dostęp do części administracyjnej aplikacji WWW.

Na poziomie katalogu użytkownikowi mogą być przypisane następujące uprawnienia:

- Dostęp - dzięki niemu użytkownik widzi katalog w drzewku biblioteki.
- Listowanie - umożliwia użytkownikowi przeglądanie zawartości katalogu (tj. publikacji i podkatalogów) oraz publikowanych edycji publikacji umieszczonych w katalogu.
- Odczyt - umożliwia użytkownikowi przeglądanie wszystkich wydań (opublikowanych i nie opublikowanych) wszystkich publikacji zamieszczonych w katalogu.
- Edycja struktury - umożliwia użytkownikowi redagowanie struktury katalogu, tj. tworzenie, przenoszenie i usuwanie podkatalogów.
- Tworzenie publikacji - umożliwia użytkownikowi tworzenie nowych publikacji w katalogu.
- Zarządzanie publikacjami - umożliwia użytkownikowi usuwanie publikacji z katalogu.
- Zarządzanie prawami - umożliwia użytkownikowi dokonywanie zmian w prawach dostępu do katalogu.

Na poziomie całej publikacji przyznane mogą być następujące trzy rodzaje praw:

- Przeglądanie - Prawo do odczytania wszystkich opublikowanych wydań publikacji.
- Odczyt - Prawo do odczytania wszystkich wydań publikacji.
- Zarządzanie - Prawo do zarządzania publikacją (np. tworzenia nowego wydania lub przyznania praw dostępu).

Każde prawo, niezależnie od tego, jakiego obiektu biblioteki dotyczy, może mieć jeden z pięciu stanów:

- Przyznane - Prawo jest przyznane bezpośrednio użytkownikowi.
- Od grupy - Użytkownik jest członkiem grupy, która ma przyznane prawo.
- Odziedziczone - Prawo zostało przyznane jednemu z obiektów nadrzędnych (np. katalogowi nadrzędnemu).
- Implikowane - Prawo jest przyznane ze względu na posiadanie innego prawa (np. użytkownik mający prawo Zarządzania, ma również implikowane prawo Czytania).
- Nie przyznane - Prawo nie jest przyznane.

Niezbędnym elementem kontroli dostępu do treści cyfrowych jest autentykacja użytkowników. W systemie dLibra możliwe są następujące sposoby autentykacji:

- W oparciu o wewnętrzną bazę danych użytkowników usługi *User Server*, poprzez podanie nazwy użytkownika i hasła lub poprzez nazwę użytkownika i adres IP komputera, z którego loguje się dany użytkownik.
- W oparciu o zewnętrzny serwer pojedynczego logowania CAS, poprzez podanie nazwy użytkownika i hasła.
- W oparciu o zewnętrzny serwer LDAP, poprzez podanie nazwy użytkownika i hasła.

W celu wykorzystania przy autoryzacji możliwości, jakie dają serwery LDAP wprowadzono w systemie dLibra mechanizm dynamicznych grup LDAP. Grupy takie definiowane są poprzez zestaw atrybutów, jakie powinien posiadać zautentykowany w serwerze LDAP użytkownik. Jeżeli użytkownik ten ma odpowiednie atrybuty, ma on takie same prawa dostępu do publikacji, jak dynamiczna grupa LDAP. Rozwiązanie to jest wykorzystywane w Kujawsko – Pomorskiej Bibliotece Cyfrowej do udostępniania skryptów akademickich i innych materiałów edukacyjnych i umożliwia łatwe grupowanie studentów ze względu na lata czy kierunki studiów. Rozwiązania tego typu byłby nieosiągalne gdyby do budowy biblioteki cyfrowej wykorzystać np. nowozelandzki Greenstone [10], jeden z najpopularniejszych tego typu systemów na świecie.

## 5. Podsumowanie i wnioski

Przedstawione w niniejszym artykule oprogramowanie umożliwia budowanie rozproszonych bibliotek cyfrowych. Jak wspomniano, wykorzystywane jest ono obecnie w kilku ośrodkach akademickich w Polsce, a liczba instalacji wciąż wzrasta. Rosnąca liczba instalacji systemu, a co za tym idzie, jego użytkowników, wymusza, aby oprogramowanie to było skalowalne i elastyczne. Dzięki takim mechanizmom, jak opisany powyżej system zarządzania metadanymi, czy mechanizmy autentykacji i autoryzacji, może być ono wykorzystywane do tak zróżnicowanych zastosowań jak budowa regionalnej cyfrowej biblioteki starodruków czy akademickiej biblioteki prac naukowych [11].

Rosnąca liczb instalacji skłania również do opracowania mechanizmów wzajemnej komunikacji pomiędzy poszczególnymi bibliotekami cyfrowymi. Mechanizmy takie są przedmiotem aktualnie prowadzonych prac w projekcie dLibra. Docelowo mają one umożliwić rozproszone wyszukiwanie oraz wymianę metadanych pomiędzy uruchomionymi w sieci PIONIER bibliotekami cyfrowymi.

Docelowo system dLibra ma stać się platformą do budowy takich bibliotek cyfrowych, które będzie można wykorzystywać w ramach zaawansowanych scenariuszy wykorzystania usług gridowych zarówno w gridach obliczeniowych jak i informacyjnych [12]. Opracowanie modelu takiej architektury jest przedmiotem dalszych prac badawczo-rozwojowych prowadzonych w PCSS.

## Bibliografia:


[1] Wielkopolska Biblioteka Cyfrowa, http://www.wbc.poznan.pl/

[2] Biblioteka Cyfrowa Politechniki Wrocławskiej, http://dlib.bg.pwr.wroc.pl/

[3] Podlaska Biblioteka Cyfrowa, http://pbc.biaman.pl/

[4] Kujawsko – Pomorska Biblioteka Cyfrowa, http://kpbc.umk.pl/

[5] Mazurek, C., Werla, M. – "Distributed Services Architecture in dLibra Digital Library Framework". 8th International Workshop of the DELOS Network of Excellence on Digital Libraries on Future Digital Library Management Systems, 29.03-01.04.2005, Schloss Dagstuhl, Germany. Workshop Proceedings.

[6] Hicks, M.; Jagannathan, S.; Kesley, R.; Moore, J.-T.; Ungureanu, C. "Transparent Communication for Distributed Objects in Java". ACM Java Grande Conference, pages 160-170, June 1999.

[7] Dublin Core Metadata Element Set wersja 1.1, http://dublincore.org/documents/dces/

[8] Lagoze, C.; Van de Sompel, H. – "The Open Archives Initiative: Building a low-barrier interoperability framework", pages 54-62, Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, VA, USA, June 2001.

[9] Hammersley , B. "Content Syndication with RSS". O'Reilly. 1st Edition. March 2003.

[10] Greenstone Digital Library User's Guide, http://prdownloads.sourceforge.net/greenstone/User-en.pdf

[11] C. Mazurek, J. A. Nikisch, M. Stroiński : Zarządzanie zdigitalizowaną biblioteką i systemy kontroli dostępu na przykładzie Wielkopolskiej Biblioteki Cyfrowej. Seminarium CPI.

[12] Kosiedowski, M.; Mazurek, C; Werla, M. – „Digital Library Grid Scenarios" in European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, 25-26.05.2004, London, U.K. Workshop Proceedings, p. 189 – 196.