Biblioteka cyfrowa jako otwarte, internetowe repozytorium publikacji

Marcin Heliński, Cezary Mazurek, Tomasz Parkoła, Marcin Werla
Poznańskie Centrum Superkomputerowo – Sieciowe
ul. Noskowskiego 12/14, 61-704 Poznań
{helin,mazurek,tparkola,mwerla}@man.poznan.pl

1. Wstęp

Tematyka bibliotek cyfrowych jest obecnie przedmiotem wielu badań naukowych toczących się w różnych krajach. W ramach piątego i szóstego programu ramowego Unii Europejskiej utrzymywana jest działalność organizacji DELOS Network of Excellence In Digital Libraries[1]. Organizacja ta ma na celu m.in. umożliwienie wymiany pomysłów i doświadczeń między osobami zajmującymi się tematyką bibliotek cyfrowych. W tym celu urządza ona tematyczne warsztaty związane z najbardziej aktualnymi problemami dotyczącymi bibliotek cyfrowych. W roku 2005 miały miejsce kolejne warsztaty DELOS, zatytułowane *„Future Digital Library Management Systems: System Architecture and Information Access”*[2] oraz *„Digital Repositories: Interoperability and Common Services”*[3]. W ramach tych warsztatów przedstawiano rezultaty aktualnie prowadzonych prac badawczych z zakresu architektury bibliotek cyfrowych, w tym prac prowadzonych w ramach programu PIONIER[4] [1][2]. Prezentowane badania skupiają się przede wszystkim na problemach rozproszonego wyszukiwania, replikacji danych czy udostępniania danych zewnętrznym systemom przy pomocy jednego wybranego protokołu.

Architektura bibliotek cyfrowych oraz współpraca z zewnętrznymi systemami była również jednym z tematów poruszanych na kolejnych konferencach *European Conference on Research and Advanced Technology for Digital Libraries* (ECDL)[5]. Na konferencjach tych omawiano między innymi sposoby dostępu do kolekcji bibliotek cyfrowych, ujednolicone interfejsy przeszukiwania oraz wykorzystanie bibliotek cyfrowych do przechowywania wyników eksperymentów naukowych. Na ostatnich konferencjach ECDL przedstawiano również konkretne projekty mające na celu ułatwienie tworzenia bibliotek cyfrowych, takie jak FEDORA[6], DSPACE[7] czy Greenstone[8]. W szeregu tych systemów umieścić można również system dLibra[9] rozwijany w ramach prac badawczo-rozwojowych prowadzonych w Poznańskim Centrum Superkomputerowo-Sieciowym[10]. System ten posłużył już do budowy sześciu bibliotek cyfrowych w Polsce.

---

[1] http://www.delos.info/
[2] http://ii.umit.at/research/delos_website/8thworkshop.html
[3] http://www.ukoln.ac.uk/events/delos-rep-workshop/
[4] http://www.pionier.gov.pl/
[5] http://www.ecdl2003.org/, http://www.ecdl2004.org/, http://www.ecdl2005.org/
[6] http://fedora.info/
[7] http://www.dspace.org/
[8] http://www.greenstone.org/
[9] http://dlibra.psnc.pl/
[10] http://www.man.poznan.pl/

2. Wielkopolska Biblioteka Cyfrowa

Pierwszą biblioteką cyfrową opartą na systemie dLibra była Wielkopolska Biblioteka Cyfrowa (WBC)[11]. Oficjalnie została ona uruchomiona w roku 2002 jako wynik współpracy Poznańskiej Fundacji Bibliotek Naukowych, Poznańskiego Centrum Superkomputerowo-Sieciowego oraz poznańskich bibliotek. Główne cele WBC to:

- zwiększenie dostępności najczęściej wykorzystywanych przez studentów podręczników i skryptów,
- zwiększenie efektywności pracy z podręcznikami akademickimi i szkolnymi,
- ułatwienie dostępu do wybranych prac naukowych (szczególnie dotyczy to monografii) naukowcom z kraju i zagranicy,
- ułatwienie, a w niektórych wypadkach wręcz umożliwienie, dostępu do źródeł informacji przechowywanych w bibliotekach i archiwach, ale ze względów bezpieczeństwa udostępnianych wyjątkowo nielicznej grupie użytkowników,
- stworzenie cyfrowych kopii najcenniejszych dzieł przechowywanych w bibliotekach i archiwach,
- obniżenie kosztów udostępniania źródeł informacji w bibliotekach.

W chwili obecnej Wielkopolska Biblioteka Cyfrowa zawiera ponad 6.000 publikacji i składa się z czterech następujących kolekcji:

- **Materiały dydaktyczne**. Zasób obejmuje skrypty, podręczniki i monografie naukowe wydane lokalnie.
- **Dziedzictwo kulturowe**. Zasób ten obejmuje najcenniejsze zabytki piśmiennictwa znajdujące się w zbiorach bibliotek poznańskich. Poza tym w jego skład wchodzą również dzieła historyczne i dzieła z zakresu literatury pięknej wydane głównie w wieku XIX (m. in. Bibljoteka Pisarzów Polskich).
- **Materiały regionalne**. Zasób obejmuje dokumenty dotyczące Poznania i Wielkopolski. Obok monografii historycznych, znajdują się tu liczne dokumenty archiwalne (akty lokacyjne, przywileje, dekrety), ulotki reklamowe poznańskich firm, katalogi wystaw, statuty poznańskich stowarzyszeń, ulotki wyborcze itp. Najstarsze materiały zgromadzone w tej kolekcji pochodzą z wieku XIII.
- **Muzykalia**. Zasób obejmuje głównie nuty ze zbiorów biblioteki Akademii Muzycznej w Poznaniu.

WBC jest w dalszym ciągu głównym czynnikiem wpływającym na kierunki rozwoju oprogramowania dLibra [3].

3. Oprogramowanie dLibra

dLibra jest zestawem oprogramowania narzędziowego wspomagającego budowę biblioteki cyfrowej oraz zarządzanie jej zawartością. Środowisko to została zaprojektowane, aby umożliwić przechowywanie i udostępnianie zróżnicowanych pod wieloma względami kolekcji obiektów cyfrowych takich jak zdigitalizowane książki (w formatach PDF, DjVu [4], HTML itp.), pliki audio oraz wideo. System dLibra oferuje swoim użytkownikom wiele interesujących cech, takich jak rozbudowane możliwości opisu bibliograficznego, mechanizm

[11] http://www.wbc.poznan.p/

wersjonowania przechowywanych treści, kompleksowa kontrola dostępu do przechowywanych zasobów, zabezpieczenia przed kopiowaniem treści czy przeszukiwanie treści i opisów publikacji [5].

Oprogramowanie dLibra to gotowy do użycia system, który może być z powodzeniem stosowany w instytucjach takich jak biblioteki publiczne czy akademickie oraz firmy wydawnicze. Obecnie dLibra jest wykorzystywana między innymi w następujących bibliotekach cyfrowych:

- Wielkopolska Biblioteka Cyfrowa (http://www.wbc.poznan.pl/)
- Dolnośląska Biblioteka Cyfrowa (http://dlib.bg.pwr.wroc.pl/)
- Kujawsko-Pomorska Biblioteka Cyfrowa (http://kpbc.umk.pl/)
- Zielonogórska Biblioteka Cyfrowa (http://zbc.uz.zgora.pl/)
- Biblioteka Cyfrowa Politechniki Łódzkiej (http://ebipol.p.lodz.pl/)
- Biblioteka Cyfrowa Uniwersytetu Wrocławskiego (http://www.bu.uni.wroc.pl/dlibra/)

Oprogramowanie dLibra jest napisane przy pomocy języka Java™ dzięki czemu nie narzuca ograniczeń na wykorzystywany system operacyjny – może to być zarówno system Windows jak i Linux, czy też MacOS. Od około roku oprogramowanie dLibra jest również niezależne od rodzaju bazy danych – do jego uruchomienia może zostać wykorzystana zarówno komercyjna baza danych Oracle jak i darmowe bazy PostgreSQL i MySQL. Dzięki wykorzystanym rozwiązaniom technicznym oprogramowanie dLibra nie sprawia również żadnych dodatkowych trudności podczas jego aktualizacji – większość czynności odbywa się w pełni automatycznie w sposób niewidoczny dla użytkownika.

Każdy zasób cyfrowy wprowadzony do biblioteki cyfrowej opartej o system dLibra może być opisany zestawem atrybutów. Opis ten może zostać wprowadzony ręcznie lub też zaimportowany z plików w formacie MARC lub RDF [6]. Wykorzystywany zestaw atrybutów jest w pełni konfigurowalny i może zostać w łatwy sposób dostosowany do potrzeb konkretnej biblioteki cyfrowej. W swojej podstawowej postaci system dLibra oferuje 15 atrybutów zgodnych ze standardem DublinCore w wersji 1.1. Wartości poszczególnych atrybutów są gromadzone w słowniku, wspierającym grupowanie wyrażeń o podobnym znaczeniu. Grupowanie to jest następnie wykorzystywane w celu poprawy wyników wyszukiwania. Ponadto każdy zasób może być przechowywany w systemie w wielu wersjach, z których każda może mieć odrębny opis i prawa odczytu.

dLibra wyróżnia trzy kategorie użytkowników biblioteki cyfrowej: czytelników, redaktorów oraz administratorów. Czytelnicy mają możliwość odczytu opisów i publikacji gromadzonych w bibliotece cyfrowej. Mogą oni przeglądać i przeszukiwać zawartość poszczególnych kolekcji. Przeszukiwanie może odbywać się przy pomocy zaawansowanych kreatorów budowy zapytań oraz indeksów wartości poszczególnych atrybutów. Dostęp czytelnika jest realizowany poprzez strony WWW biblioteki cyfrowej generowane przez dLibrę.

Redaktorzy i administratorzy systemu dLibra mają do swojej dyspozycji odrębną aplikację. Redaktorzy mogą dodawać, edytować i usuwać zasoby gromadzone w bibliotece cyfrowej. Mogą oni również opisywać te zasoby metadanymi, organizować je w kolekcje oraz ustalać zasady dostępu dla określonych użytkowników oraz grup użytkowników. Administratorzy biblioteki cyfrowej mogą zarządzać użytkownikami, grupami użytkowników, katalogami, kolekcjami i atrybutami używanymi przy opisie.

4. Dostęp do informacji w Internecie

Wraz z gwałtownym wzrostem liczby zasobów dostępnych w Internecie, bardzo wzrósł stopień skomplikowania wyszukiwania pożądanych czy udostępnianych przez nas informacji. Olbrzymi rozmiar i zróżnicowanie danych w Internecie, zarówno pod względem formy czy zawartości jaki jakości, doprowadziły do tego, iż osoby poszukujące informacji korzystają przede wszystkim z wyszukiwarek internetowych takich jak na przykład Google. Z tego też powodu bardzo istotne jest, aby kontrolować i polepszać widoczność udostępnianych przez nas zasobów w tego typu systemach. Widoczność tą można na przykład sprawdzić poprzez zapytanie site:adres.naszej.strony wydane wyszukiwarce Google. W odpowiedzi otrzymamy przybliżoną listę oraz liczbę stron, które dostępne są dla internautów postrzegających zasoby internetowe przede wszystkim przez pryzmat wyszukiwarek. Zapytanie takie wydane dla kilku bibliotek cyfrowych w dniu 20.11.2005 r. dało następujące wyniki:

- Większe biblioteki:
  - PBI[12] – 26 243 publikacje – zapytanie *site:pbi.edu.pl* około 74 500 wyników (na jedną publikację przypadło średnio około 2,8 wyniku)
  - WBC – 6 323 publikacje – zapytanie *site:www.wbc.poznan.pl* dało około 40 100 wyników (na jedną publikację przypadło średnio około 6,3 wyniku)
  - KPBC – 1 000 publikacji – zapytanie *site:kpbc.umk.pl* dało około 19 900 wyników (na jedną publikację przypadło średnio około 19,9 wyniku)
- Mniejsze biblioteki
  - WBSS PG[13] – 82 publikacje (wartość przybliżona) – zapytanie *site:www.wbss.pg.gda.pl dało* 852 wyniki (na jedną publikację przypadło średnio około 10,3 wyniku)
  - DBC – 133 publikacje – zapytanie *site:dlib.bg.pwr.wroc.pl* dało około 21 000 wyników (na jedną publikację przypadło średnio około 157,8 wyniku)

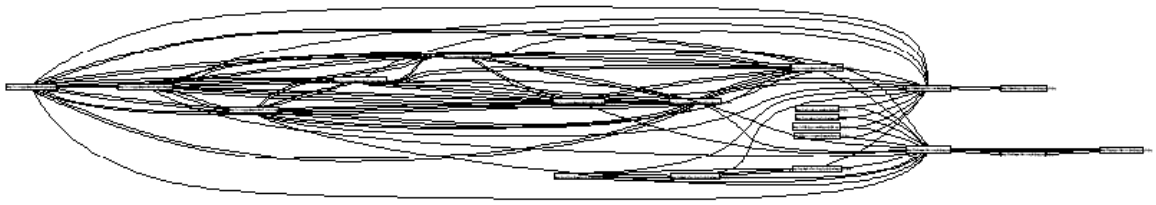
Dane te pokazują, jak zróżnicowana może być widoczność udostępnianych zasobów cyfrowych w zależności od przyjętych rozwiązań technicznych i organizacyjnych. Komentarza wymaga tutaj kwestia rozdzielenia bibliotek cyfrowych na mniejsze i większe. Zostało ono dokonane, gdyż rozróżnienie takie wprowadzają same wyszukiwarki internetowe. Indeksują one tylko pewną liczbę stron z danej biblioteki. Im mniejsze jest indeksowana biblioteka tym korzystniej wypada stosunek stron zaindeksowanych do niezaindeksowanych. Bardzo znamienne są wyniki wyświetlane dla Polskiej Biblioteki Internetowej. Mimo iż, liczba wyników jest dość duża, to wyniki te niestety można by sprowadzić do kilku czy kilkunastu stron zawierających podstawowe informacje na temat PBI. Obecny sposób udostępniania treści cyfrowych w PBI powoduje iż treści te nie są dostępne dla wyszukiwarek internetowych, a zamiast tego indeksowane są identyczne kopie kilku informacyjnych stron biblioteki. Przez to zawartość największej obecnie polskiej biblioteki cyfrowej jest praktycznie niewidoczna w wyszukiwarce Google.

5. Rozproszone biblioteki cyfrowe w sieci PIONIER

Zwiększenie widoczności zasobów cyfrowych w Internecie można osiągnąć wykorzystując możliwości jakie daje wzrost liczby bibliotek cyfrowych dostępnych w polskim Internecie. Obecnie każda z bibliotek cyfrowych stanowi odrębny punkt dostępu do

[12] http://www.pbi.edu.pl/
[13] http://www.wbss.pg.gda.pl/

zasobów gromadzonych w tej właśnie bibliotece cyfrowej. Dzięki połączeniu bibliotek cyfrowych w rozproszoną platformę możliwe jest, aby każda z bibliotek cyfrowych stanowiła punkt dostępu do zasobów zgromadzonych w ramach całej rozproszonej platformy. System dLibra w wersji 2.2 umożliwia stworzenie takiej właśnie platformy dzięki wykorzystaniu mechanizmów rozproszonego pobierania i indeksowania metadanych opartego na protokole OAI-PMH [7]. Wykorzystanie protokołu OAI-PMH umożliwia włączenie do platformy bibliotek cyfrowych również bibliotek cyfrowych nie wykorzystujących oprogramowania dLibra, a jedynie udostępniających interfejs zgodny z najnowszą wersją protokołu OAI-PMH. Podobne rozwiązania są realizowane w innych krajach (por. Rys. 1), jednak żadne z wykorzystywanych obecnie narzędzi do budowy bibliotek cyfrowych nie pozwala na łatwy dostęp do zasobów rozproszonych z każdej biblioteki należącej do danej platformy bibliotek cyfrowych. W tym zakresie platforma bibliotek cyfrowych utworzona w sieci PIONIER z wykorzystaniem oprogramowania dLibra będzie rozwiązaniem nowatorskim na skalę międzynarodową.



Rys 1. Schemat sieci rozproszonych bibliotek cyfrowych w Niemczech
(z dnia 20.11.2005 za OAI Registry at UIUC[14])

Proponowane przez dLibrę rozwiązanie oparte jest na okresowym (np. raz na dobę) odpytywaniu poszczególnych bibliotek cyfrowych o zmiany w metadanych dotyczących obiektów przechowywanych w danej bibliotece cyfrowej. Zmiany te są pobieranie, a następnie wykonywane jest uaktualnienie i indeksacja tych metadanych. Dzięki indeksacji metadanych dotyczących rozproszonych zasobów, przeprowadzanej w każdej z bibliotek cyfrowych, możliwe jest udostępnianie użytkownikom tych bibliotek cyfrowych funkcji przeszukiwania metadanych dotyczących wszystkich zasobów dostępnych w ramach danej platformy. Rozwiązanie takie nie powoduje również wzrostu obciążenia poszczególnych bibliotek cyfrowych poprzez zapytania przychodzące z zewnętrznych bibliotek cyfrowych. Każda z bibliotek cyfrowych ma za zadanie obsłużyć wyszukiwania tylko swoich czytelników oraz raz na pewien czas (np. raz na dobę) udostępnić innym bibliotekom cyfrowym informacje o zmianach.

Kolejnym etapem rozwoju rozproszonej platformy bibliotek cyfrowych realizowanej w sieci PIONIER z wykorzystaniem protokołu OAI-PMH oraz oprogramowania dLibra będzie udostępnienie funkcji pozwalających na przeszukiwanie treści rozproszonych zasobów. Rozwiązanie takie będzie wiązało się z opracowaniem mechanizmu pobierania i indeksowania zasobów gromadzonych w rozproszonych bibliotekach cyfrowych, a następnie z udostępnieniem tego mechanizmu w ramach już istniejących bibliotek cyfrowych.

Pełne wykorzystanie możliwości platformy rozproszonych bibliotek cyfrowych będzie się wiązało z koniecznością wykonania szeregu prac organizacyjnych skupiających się na opracowaniu standardów dotyczących m.in. schematów metadanych używanych do opisu zasobów cyfrowych. Początkowo do tego celu może być wykorzystywany np. standard

[14] http://gita.grainger.uiuc.edu/registry/FriendsGraph.asp?type=cmap

DublinCore, jednak nawet on wymaga pewnej interpretacji jeżeli chodzi o przygotowywanie opisów zróżnicowanych materiałów cyfrowych, takich jak np. zeskanowane pocztówki, partytury, starodruki czy artykuły naukowe i rozprawy doktorskie. Kolejnym zadaniem będzie opracowanie ustaleń umożliwiających tworzenie wirtualnych, rozproszonych kolekcji czy wystaw gromadzących obiekty cyfrowe pochodzące z poszczególnych bibliotek rozproszonej platformy.

6. Zakończenie

Usługi rozproszonej biblioteki cyfrowej powinny być dostępne nie tylko poprzez specyficzne dla konkretnych usług interfejsy, ale również poprzez protokoły będące uznanymi w danym obszarze standardami. Przykładem jest tu protokół OAI-PMH czy protokół WebDAV używany do dostępu do wersjonowanej treści cyfrowej. Interesujące w tym kontekście mogą być również inne protokoły, jak np. związane z sieciami P2P *(ang. Peer to Peer).* Dzięki takiemu podejściu możliwe będzie wykorzystanie usług rozproszonych bibliotek cyfrowych w wielu istniejących już systemach takich jak na przykład systemy e-learningowe, portale edukacyjne, systemy dostarczania treści multimedialnej itp. Oznacza to, iż rozproszone biblioteki cyfrowe przy odpowiedniej strukturze i oferowanych usługach zdynamizują powstawanie nowych usług i aplikacji funkcjonujących w oparciu o dobre i pełne zasoby informacyjne.

Bibliograf

[1] Mazurek C., Werla M. “Distributed Services Architecture in dLibra Digital Library Framework” in Proceedings of the 8th International Workshop of the DELOS Network of Excellence on Digital Libraries on Future Digital Library Management Systems, Schloss Dagstuhl, Germany, 2005.

[2] Mazurek C., Werla M. “Digital Object Lifecycle in dLibra Digital Library Framework” in Proceedings of the 9th International Workshop of the DELOS Network of Excellence on Digital Libraries on Digital Repositories, Heraklion, Crete, 2005.

[3] Mazurek C., Nikisch J.A., Werla M. „Rozwój Wielkopolskiej Biblioteki Cyfrowej, a zmiany funkcjonalności systemu dLibra”, Materiały konferencyjne - Konferencja „Digitalizacja Zbiorów Bibliotecznych”, Warszawa, 2005.

[4] “DjVu: A Tutorial”, http://www.djvuzone.org/support/tutorial/index.html

[5] Mazurek C., Stroiński M., Werla M. „Wdrażanie regionalnych bibliotek cyfrowych w sieci PIONIER w oparciu o środowisko dLibra”. Materiały konferencyjne z IV Krajowej Konferencji Naukowej INFOBAZY 2005. Bazy danych dla nauki, Gdańsk, 25 - 27.09.2005,

[6] Klyne, Graham; Caroll, Jeremy J. – “Resource Description Framework (RDF): Concepts and Abstract Syntax”, http://www.w3.org/TR/rdf-concepts/

[7] Lagoze, Carl; Van de Sompel, Herbert – “The Open Archives Initiative Protocol for Metadata Harvesting”, http://www.openarchives.org/OAI/openarchivesprotocol.html