POLSKIE BIBLIOTEKI CYFROWE 2010 • **str. 17-21**

Biblioteki cyfrowe – główne kierunki rozwoju

CEZARY MAZUREK, MACIEJ STROIŃSKI, JAN WĘGLARZ

Poznańskie Centrum Superkomputerowo-Sieciowe

e-mail: {mazurek|stroins|weglarz}@man.poznan.pl

Streszczenie

W artykule przedstawione zostały nowe obszary zastosowań dla infrastruktury bibliotek cyfrowych, związane głównie z zastosowaniem nowych mechanizmów i usług w warsztacie pracy naukowców. Infrastruktura bibliotek cyfrowych z jej rozbudowanym zasobem metadanych jest jednym z głównych czynników stymulujących rozwój w jej otoczeniu, zintegrowanych systemów wiedzy. Koncepcja takiego systemu wraz z repozytorium danych źródłowych stanowi przedmiot prac badawczo-rozwojowych prowadzonych w ramach projektu SYNAT. W artykule przedstawiono główne elementy tej koncepcji oraz założenia co do funkcji realizowanego w projekcie Zintegrowanego Systemu Wiedzy. Ponadto, nowy kierunek rozwoju środowisk bibliotek cyfrowych powiązany jest z archiwizacją procesów badawczych. Prace badawcze, realizowane w ramach projektu Wf4Ever polegają na opracowaniu koncepcji zachowania procesu badawczego wraz ze skojarzonymi z nim obiektami w środowisku bibliotek cyfrowych. Oba przedstawione w artykule główne kierunki rozwoju bibliotek cyfrowych powiązane są z realizacją prototypowych implementacji systemów przeznaczonych do wsparcia nauk humanistycznych oraz astronomii i bioinformatyki.

Słowa kluczowe: biblioteki cyfrowe, archiwizacja procesów badawczych, *workflow*, zintegrowane systemy wiedzy

Wstęp

Dynamiczny rozwój sieci bibliotek cyfrowych w Polsce stworzył dogodne warunki dla istotnego wzrostu ilości zasobów udostępnianych przez polskie biblioteki, archiwa, muzea jak również prywatne osoby w ramach infrastruktury informatycznej nauki [1]. Ponad 300 instytucji pracujących obecnie w ramach tej sieci, od 2002 roku przetworzyło do postaci cyfrowej i udostępniło w Internecie blisko pół miliona obiektów. Dzięki uruchomionej w 2007 roku Federacji Bibliotek Cyfrowych, zasoby te są jednocześnie widoczne w Europejskiej Bibliotece Cyfrowej Europeana [2]. Wraz z tymi osiągnięciami, pojawiają się nowe wyzwania związane z szerszym wykorzystaniem i popularyzacją bibliotek cyfrowych. Dotyczą one kilku zasadniczych kierunków w rozwoju istniejącej infrastruktury bibliotek cyfrowych, z których wiodącym jest opracowanie nowych mechanizmów pozyskiwania wiedzy z zasobów takich agregatorów jak Federacja Bibliotek Cyfrowych. Poprzez rozwój takich mechanizmów, biblioteki cyfrowe mogą stać się ważnym składnikiem zintegrowanych platform wiedzy umożliwiających współpracę i prowadzenie badań w oparciu o otwarty dostęp do danych.

Pierwszym krokiem w kierunku budowy zintegrowanych systemów wiedzy było uruchomienie w 2007 roku Federacji Bibliotek Cyfrowych w sieci PIONIER [3]. Pozwoliło to na różnokontekstową agregację metadanych i informacji o strukturze bibliotek cyfrowych na poziomie krajowym. W ten sposób, do istniejących już elementów tej infrastruktury, takich jak sieć światłowodowa Polskiego Internetu Optycznego oraz zasoby obliczeniowe zorganizowane w strukturze gridów, dodana została warstwa danych naukowych oraz pozyskiwania informacji z istniejących zasobów bibliotek cyfrowych. Taka struktura umożliwiła realizację nowych wyzwań tak w zakresie nauk technicznych jak i humanistycznych. W kolejnych rozdziałach omówiono przykłady dalszego rozwoju polskiej sieci bibliotek cyfrowych pod kątem budowy zintegrowanych systemów wiedzy oraz archiwizacji procesów badawczych.

Zintegrowane systemy wiedzy

Jednym z najistotniejszych zadań, jakie stoją przed twórcami narzędzi dla budowy bibliotek cyfrowych w kontekście zastosowań naukowych, jest rozwój usług i aplikacji, które pozwolą na wprowadzenie nowego wymiaru ba-

dań naukowych w dziedzinie humanistyki. Podstawowy składnik tego warsztatu naukowego to przechowanie i łatwy dostęp do unikalnych dokumentów (czyli zasobów informacyjnych). Rzeczywista korzyść pojawia się jednak dopiero wtedy, gdy z określonym obiektem (zasobem) będą mogły być powiązane (najlepiej automatycznie) inne zasoby lub też procesy, umożliwiające skuteczne jego przetwarzanie w celu odkrycia przesłanek dla nowych zależności i zbudowania nowych zasobów wiedzy. Przy czym, istotne jest w tym wypadku wprowadzenie określonych struktur pośredniczących oraz mechanizmów definiowania ograniczeń i warunków ukończenia tego typu procesów z jak najlepszym skutkiem.

Wykorzystanie tak ogólnego i elastycznego schematu metadanych jak Dublin Core pozwala serwisom agregującym (jak np. FBC) na przyłączanie usług sieciowych udostępniających informacje z wielu różnych dziedzin [4]. Niestety takie podejście ma też słabe strony. Zbyt ogólne definicje poszczególnych elementów schematu Dublin Core i brak szeroko przyjętych ogólnoświatowych wytycznych w zakresie interpretacji tych definicji w poszczególnych domenach bardzo często uniemożliwiają zaawansowaną integrację informacji agregowanych za pomocą protokołu OAI-PMH.

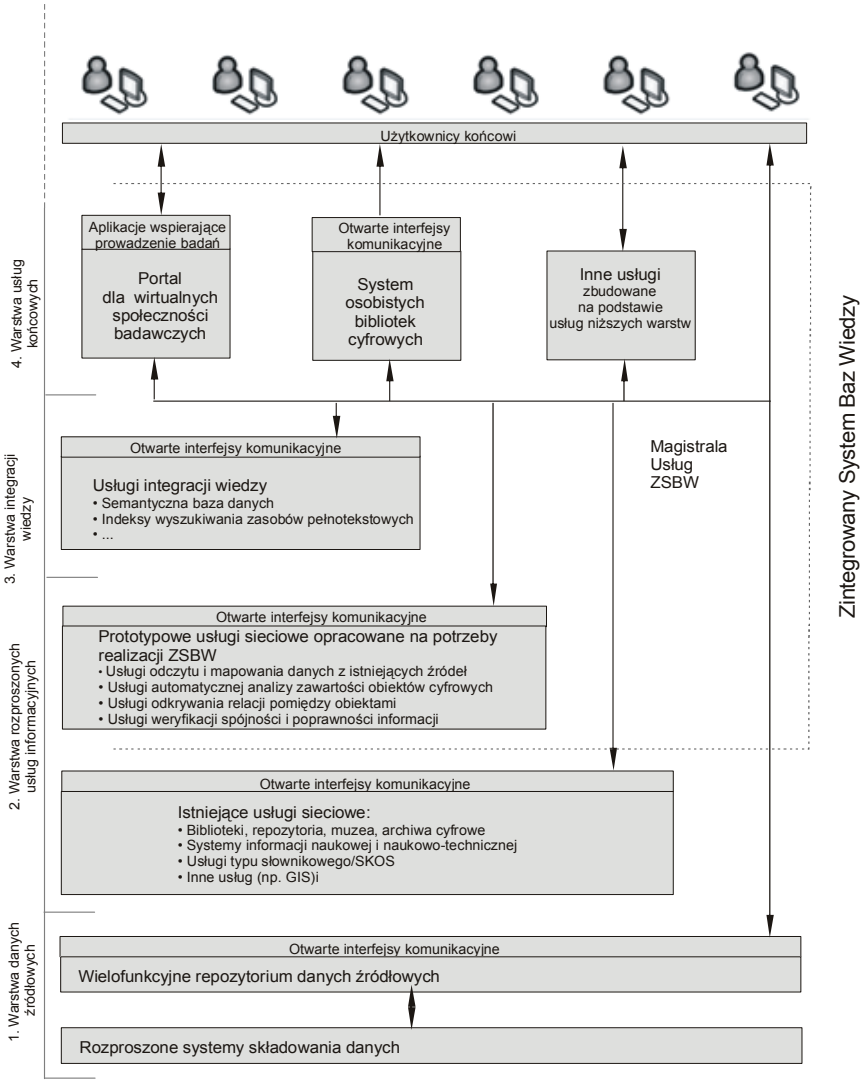
W związku z tym projekty pragnące osiągnąć większy poziom integracji informacji niż tylko wspólne indeksy poszczególnych elementów schematu Dublin Core tworzone na podstawie rozproszonych baz informacji (np. indeks tytułów czy indeks twórców), udostępniają precyzyjne wytyczne dotyczące interpretacji pól Dublin Core lub nawet rekomendują wykorzystywanie rozszerzeń Dublin Core (tzw. kwalifikatorów) w celu doprecyzowania znaczenia przekazywanych za pomocą protokołu OAI-PMH informacji. Przykładami mogą być tutaj wytyczne projektu DRIVER [5], agregującego w swoim portalu efekty prac naukowych z całej Europy oraz schemat Europeana Semantic Elements [6] wraz z wytycznymi będący podstawą do agregacji kilku milionów obiektów europejskiego dziedzictwa kulturowego w portalu Europeana. Takie podejście, połączone z wykorzystaniem technologii semantycznych, może być podstawą do realizacji nowatorskich interfejsów użytkownika i metod eksploracji informacji, takich jak prototypowe Laboratorium Myśli (ang. *Tought Lab*) dostępne w portalu Europeana (http://www.europeana.eu/portal/thought-lab.html). Skonstruowanie takiego laboratorium wymaga, poza użyciem do reprezentacji informacji jednego schematu metadanych, ściśle ustandaryzowanego słownictwa i korzystania z kontrolowanych słowników tam gdzie tylko jest to możliwe.

Realizacja zintegrowanego systemu wiedzy wymaga wieloaspektowej integracji i interoperacyjności istniejących źródeł wiedzy oraz narzędzi i usług sieciowych. Próbę realizacji takiego systemu podjęto w ramach projektu SYNAT, a konkretnie zadania badawczego obejmującego realizację, m.in. Wielofunkcyjnego Repozytorium Danych Źródłowych oraz Zintegrowanego Systemu Wiedzy (ZSW).

Punktem wyjścia do realizacji systemu jest czteropoziomowa architektura danych i usług obejmująca następujące warstwy (ryc. 1).

1) Warstwa danych źródłowych, służąca długoterminowemu przechowaniu danych źródłowych dla nauki, składająca się z:
   - różnorodnych systemów składowania danych;
   - prototypowych usług opracowanych w trakcie realizacji Wielofunkcyjnego Repozytorium Danych Źródłowych.
2) Warstwa rozproszonych usług informacyjnych służąca pozyskiwaniu i przetwarzaniu wiedzy z rozproszonych źródeł danych, składająca się z:
   - istniejących usług sieciowych takich jak biblioteki, muzea i archiwa cyfrowe czy systemy informacji naukowej;
   - prototypowych usług opracowanych w trakcie realizacji ZSW, w szczególności usług agregowania i przetwarzania informacji, niezbędnych w procesie integracji wiedzy w warstwie 3.
3) Warstwa integracji wiedzy, ukierunkowana na eksploatację usług informacyjnych i eksplorację rozproszonych zasobów, składająca się z:
   - pilotażowej realizacji semantycznej bazy wiedzy;
   - prototypowych usług integrujących wiedzę.
4) Warstwa usług końcowych zapewniająca dostęp użytkownika końcowego do usług systemu zgromadzonych we wszystkich jego warstwach, składająca się m.in. z:

- portalu dla wirtualnych społeczności badawczych;
- systemu osobistych bibliotek cyfrowych dla naukowców;
- innych usług działających na podstawie funkcjonalności i wiedzy dostarczanej przez ZSW.



Ryc. 1. Architektura Zintegrowanego Systemu Wiedzy

Podstawowym zadaniem warstwy danych źródłowych jest bezpieczne i długoterminowe przechowywanie danych źródłowych, takich jak wysokiej jakości cyfrowe kopie fizycznych obiektów (tzw. kopie master, związane zazwyczaj z cyfrowym zabezpieczaniem dziedzictwa kulturowego), czy dane pochodzące z eksperymentów i pomiarów, powiązane z pracami naukowymi, często występujące zarówno w formie podstawowej, jak i wstępnie przetworzonej.

Głównymi elementami warstwy rozproszonych usług informacyjnych będą:

- biblioteki, repozytoria, muzea i archiwa cyfrowe, zawierające metadane oraz postać prezentacyjną obiektów składowanych w warstwie danych źródłowych oraz inne obiekty cyfrowe i metadane, często w różny sposób powiązane z danymi źródłowymi;

- systemy informacji naukowej i naukowo-technicznej, takie jak systemy katalogowe dla bibliotek czy systemy dokumentacji zabytków dla muzeów, a także repozytoria czasopism i artykułów naukowych, bibliografii itp.;
- usługi sieciowe dające dostęp do danych typu słownikowego, tezaurusów, taksonomii itp.;
- inne usługi udostępniające dane istotne w kontekście ZSW (np. usługi GIS, ale również odpowiednio wiarygodne encyklopedie on-line czy portale edukacyjne).

Warstwa integracji wiedzy odpowiedzialna będzie za udostępnienie banku wiedzy integrującego w jednorodnej przestrzeni informacyjnej wszystkie dane naukowe agregowane bądź wytwarzane automatycznie (np. wskutek analizy treści obiektów cyfrowych) przez rozproszone usługi informacyjne. Wraz z bankiem wiedzy udostępnione zostaną również niezbędne powiązane usługi pomocnicze, takie jak np. indeksy wyszukiwawcze zasobów pełnotekstowych. Podstawowym elementem tej warstwy będzie system reprezentacji wiedzy oparty o technologie związane z maszynowym przetwarzaniem zorganizowanej wiedzy, takie jak OWL, RDF, SKOS, a podstawą dla konstrukcji szkieletu reprezentacji wiedzy może być standard ISO 21127: 2006 (CIDOC CRM) [7].

Usługi warstwy integracji wiedzy, podobnie jak usługi niższych warstw budowanej infrastruktury, dostępne będą poprzez otwarte, zorientowane na zasoby interfejsy komunikacyjne. Interfejsy te zostaną wykorzystane, m.in. w warstwie usług końcowych do przygotowania pilotażowej wersji aplikacji dających użytkownikom dostęp do zasobów i funkcji udostępnianych przez ZSW oraz inne elementy systemu.

Poza nakreślonym w ten sposób, głównym nurtem rozwoju infrastruktury bibliotek cyfrowych zmierzającym w kierunku wsparcia w budowie zintegrowanych systemów wiedzy, nowym zastosowaniem dla dostępnej już infrastruktury jest archiwizacja procesów badawczych (ang. *Workflow*), jaka będzie realizowana w ramach projektu Wf4Ever. Obszarem zastosowań dla opracowanej tam technologii będzie astronomia oraz bioinformatyka.

Archiwizacja procesów badawczych

Nowe zastosowania dla infrastruktury bibliotek cyfrowych pojawiają się na gruncie prac badawczych prowadzonych w projekcie Wf4Ever. Celem projektu jest opracowanie oryginalnego modelu i wykonanie mechanizmów wspomagających automatyczną archiwizację kompletnych procesów badawczych (ang. *scientific workflows*), prowadzonych w środowisku eInfrastruktury i zdefiniowanych w oparciu o duże ilości danych będące rezultatem prowadzonych eksperymentów naukowych. Wyniki prac badawczych w projekcie pozwolą na:

- wykorzystanie złożonych obiektów cyfrowych zawierających dane statyczne i dynamiczne, takie jak modele przepływów pracy, historia ich wykonań oraz powiązania z innymi zasobami,
- zapewnienie możliwości dostępu, modyfikacji, współdzielenia i wielokrotnego wykorzystywania złożonych obiektów cyfrowych,
- zarządzanie cyklem życia procesów badawczych oraz powiązanych zasobów.

W ramach projektu Wf4Ever zostaną opracowane mechanizmy i usługi infrastruktury badawczej pozwalające na zachowywanie i efektywne wykorzystywanie zapisu procesów badawczych w różnych dziedzinach nauki.

Wf4Ever daje naukowcom zupełnie nowe możliwości w dziedzinie zachowywania wiedzy naukowej poprzez innowacyjną koncepcję Obiektów Badawczych opartych o przepływy pracy, która uwzględnia istotną rolę, jaką przepływy pracy odgrywają obecnie w świecie zinformatyzowanej nauki. Fundamentalnym założeniem projektu jest fakt, że przepływy pracy są zmienne w czasie i podlegają ciągłej ewolucji, często nawet poza kontrolą ich twórcy, a zapewnienie ich spójności i aktualności jest wymaganiem, które musi zostać spełnione.

Do realizacji powyższej koncepcji wykorzystane zostaną komponenty oprogramowania dLibra, umożliwiające przechowanie procesu badawczego oraz powiązanych z nim obiektów w środowisku biblioteki cyfrowej. Stworzenie referencyjnej implementacji Wf4Ever zakłada integrację podstawowych systemów informatycznych umożliwiających współdzielenie naukowych przepływów pracy i archiwizowanie obiektów cyfrowych z aplikacjami i usługami. Referencyjna implementacja będzie oparta o otwarte standardy i protokoły i będzie rozszerzać istniejące biblioteki cyfrowe o możliwości przechowywania, indeksowania oraz udostępniania zapisu procesów badawczych tak, aby otrzymać wyspecjalizowaną bibliotekę cyfrową dedykowaną przepływom pracy.

Wf4Ever połączy i rozszerzy dotychczasowe osiągnięcia partnerów projektu (iSOCO, University of Manchester, Universidad Politécnica de Madrid, Poznańskie Centrum Superkomputerowo-Sieciowe, University of Oxford, Instituto de Astrofísica de Andalucía oraz Leiden University Medical Centre) w obszarach zarządzania przepływami pracy, bibliotek cyfrowych, sieci społecznościowych oraz sieci semantycznych. Rezultaty projektu będą zweryfikowane poprzez ich wykorzystanie przez środowiska astronomów oraz biologów, zarówno do tworzenia i zarządzania zupełnie nowymi przepływami pracy, jak i do rozszerzenia możliwości współdzielenia przepływów już istniejących. Poprzez wybranie właśnie tych dwóch dziedzin zastosowań, projekt umożliwi nie tylko stworzenie nowych metod i narzędzi do zachowania wiedzy naukowej, ale także zastosowanie tych innowacji w najbardziej do tego odpowiednich dziedzinach nauki

Podsumowanie

Przedstawione w artykule główne kierunki rozwoju infrastruktury bibliotek cyfrowych w sieci PIONIER, mają umocowanie w europejskich trendach rozwoju eInfrastruktury i Internetu Przyszłości. Opracowanie pt. „Future Media Internet Challenges and the Road Ahead" pokazuje rozwój treści dostępnych w Internecie od nieustrukturyzowanych zasobów, zwykle podlegających „ręcznemu" przetwarzaniu w kierunku inteligentych treści, przechowywanych w modelu obiektowym, które potrafią się same organizować w złożone struktury. Zasoby Federacji Bibliotek Cyfrowych stanowią bardzo dobry punkt wyjścia do zintegrowanych systemów wiedzy, a środowiska bibliotek cyfrowych oparte na oprogramowaniu Libra, pozwalają na archiwizację procesów badawczych i związanych z nimi obiektów. Każdy z tych obszarów stanowi bardzo istotny czynnik w rozwoju infrastruktury badawczej niezbędnej do podejmowania nowych wyzwań w zakresie nauk humanistycznych jak i ścisłych.

Piśmiennictwo

[1] Mazurek, C., Stroiński, M., Werla, M., Węglarz, J. *Infrastruktura bibliotek cyfrowych w sieci PIONIER*. [W:] *Materiały konferencyjne*. Konferencja „Polskie Biblioteki Cyfrowe 2008", Poznań, str. 9-14. ISBN 978-83-7314-143-8.

[2] Werla M. *Polskie biblioteki cyfrowe, FBC i Europeana – etapy i bariery w przepływie informacji*. [W:] Biuletyn EBIB [Dokument elektroniczny] / red. naczelny Bożena Bednarek-Michalska, Nr 1/2010 (110) luty. Czasopismo elektroniczne, [Warszawa]: Stowarzyszenie Bibliotekarzy Polskich KWE, 2010. Tryb dostępu: http://www.ebib.info/2010/110/a.php?werla. – Tyt. z pierwszego ekranu. ISSN 1507-7187

[3] Lewandowska A., Mazurek C., Werla M. *Federacja Bibliotek Cyfrowych w sieci PIONIER – Dostęp do otwartych bibliotek cyfrowych i repozytoriów*. Warszawa: SBP KWE, 2007. ISBN 83-921757-6-X. IV Ogólnopolska Konferencja EBIB Internet w bibliotekach – Open Access, Toruń, 7-8 grudnia, 2007.

[4] Lewandowska, A. Mazurek, C.,Werla, M. *Enrichment of European Digital Resources by Federating Regional Digital Libraries in Poland.* [In:] 12th European Conference, ECDL 2008, Aarhus, Denmark, September 14-19, 2008. Proceedings Series: LNCS, Vol. 5173, pp. 256–259 (2008).

[5] Lossau, N. and Peters, D. (2008). *DRIVER: Building a Sustainable Infrastructure of European Scientific Repositories*. Liber Quarterly 18(3/4): 437-448.

[6] Europeana Semantic Elements specifications. Version 3.2.2, 18/01/2010. [on-line]. [Dostęp 30 września 2010]. Dostępny w World Wide Web: http://version1.europeana.eu/c/document_library/get_file?uuid=c56f82a4-8191-42fa-9379-4d5ff8c4ff75&groupId=10602

[7] Crofts, N., Doerr, M., Gill, T., Stead, S., and Stiff, M. *Definition of the CIDOC Conceptual Reference Model, 5.0.2 edition*, June 2005. http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.2.pdf (2005).