POZNAŃSKIE CENTRUM SUPERKOMPUTEROWO SIECIOWE

POZNAŃSKIE
CENTRUM
SUPERKOMPUTEROWO
SIECIOWE

Metadane dokumentów w bibliotekach cyfrowych

Marcin Werla, PCSS

Metadane - krótko o historii…

- Wielkopolska Biblioteka Cyfrowa rozpoczęła swoją publiczną działalność w październiku 2002 roku
  - Była to pierwsza biblioteka cyfrowa oparta na systemie dLibra
  - Preinstalowanym schematem metadanych w tym systemie był schemat Dublin Core (wtedy i przez kilka następnych lat)
- Od 2004 roku zaczęły powstawać kolejne biblioteki cyfrowe oparte na dLibrze
  - Obecnie jest ich około 60, co stanowi mniej więcej 90% bibliotek cyfrowych w Polsce
  - Każda z nich zaczynała z preinstalowanym schematem Dublin Core

Metadane - krótko o historii…

- Jak pokazały analizy prezentowane na II Konferencji „Polskie Biblioteki Cyfrowe” (9.12.2009 r.)

  The Dublin Core Metadata Element Set, Ver. 1.1 a potrzeby i oczekiwania bibliotekarzy cyfrowych - analiza przypadków - Joanna Potęga (Biblioteka Narodowa), Agnieszka Wróbel (Biblioteka Uniwersytecka w Warszawie) – http://dl.psnc.pl/biblioteka/dlibra/publication/253/content

  większość polskich bibliotek cyfrowych zdecydowała się w ostatnich latach na modyfikację domyślnego schematu poprzez:
  - dodawanie uszczegółowień pól DC
  - dodawanie nowych elementów schematu równorzędnych z polami DC
- Równocześnie brakowało wytycznych interpretacji schematu Dublin Core
  - Opisy pól schematu zawarte w standardzie były i są bardzo ogólne
  - Powstające polskie tłumaczenia zawierały błędy np.:

    „Absence of DC.RIGHTS in a record does not imply that the resource is not protected.” ➔

    „Jeżeli element Prawa nie występuje, nie można zakładać, że istnieją jakiekolwiek prawa dotyczące źródła.”

Metadane - krótko o historii…

- Wiosną 2007 roku udostępniony został „ePoradnik redaktora zasobów cyfrowych” czyli „Interpretacja schematu Dublin Core wraz z materiałami pomocniczymi dla redaktorów zasobów cyfrowych Biblioteki Cyfrowej Uniwersytetu Wrocławskiego”
  - http://www.bibliotekacyfrowa.pl/publication/14396
- ePoradnik dla każdego z elementów DC podawał jakie dane powinny być w tym elemencie zawarte oraz w jaki sposób powinny być zapisane
- Dla każdego elementu DC podano również powiązane pole (pola) schematu MARC
- Dokument zawiera też kilka załączników np. wykaz typów dokumentów czy skrótów dla atrybutu Język
- Udostępnienie ePoradnika było dużym „społecznościowym” krokiem w kierunku poprawy jakości opisów w polskich bibliotekach cyfrowych

Metadane - krótko o historii…

- W czerwcu 2007 roku udostępniona publicznie została Federacja Bibliotek Cyfrowych – usługa sieci naukowej PIONIER agregująca informacje o obiektach udostępnianych przez polskie biblioteki cyfrowe
  - http://fbc.pionier.net.pl/
- Na potrzeby kopiowania metadanych do FBC wykorzystano protokół OAI-PMH zaimplementowanego w dLibrze
  - Schemat ten wymusza wykorzystanie schematu Dublin Core
  - W związku z tym podstawowym schematem metadanych FBC również został Dublin Core
- Zaletą tego rozwiązania było łatwe osiągnięcie sposobu wymiany informacji pomiędzy bibliotekami cyfrowymi – zarówno na poziomie protokołu transmisji danych, jak i na poziomie schematu metadanych
- Wadą było uproszczenie danych w stosunku do potencjalnie bogatszych schematów stosowanych w poszczególnych bibliotekach
- Gromadzenie danych w FBC pokazało też problem rozbieżności interpretacji oraz sposobów zapisu poszczególnych pól schematu DC
  - http://dl.psnc.pl/biblioteka/dlibra/publication/210/content

Metadane - krótko o historii…

- W dniach 7-8 września 2009 r. w Gnieźnie odbyły się warsztaty na temat opracowania zasobów bibliotek cyfrowych zorganizowane przez Poznańską Fundację Bibliotek Naukowych (http://www.pfsl.poznan.pl/pbc_warsztaty)
  - W ramach warsztatów powołano 3 zespoły robocze:
    - ds. metadanych w bibliotekach cyfrowych
    - ds. synonimów
    - ds. zagadnień prawnych
- Pierwszy z zespołów rozpoczął prace nad stworzeniem nowego schematu metadanych, który można by zastosować w większości polskich bibliotek cyfrowych
  - Wiki: http://dlibra.psnc.pl/community/display/MET/
  - Dyskusja była też prowadzona na forum Biblioteka 2.0
    - http://forum.biblioteka20.pl/viewforum.php?f=12

Metadane - krótko o historii…

- W połowie 2008 r. PCSS jako operator Federacji Bibliotek Cyfrowych rozpoczął współpracę z Europeaną
- Pod koniec 2009 roku miał miejsce pierwszy transfer danych z FBC do Europeany
  - http://europeana.eu/portal/brief-doc.html?query=*:*&qf=PROVIDER:Federacja%20Bibliotek%20Cyfrowych
- W następnych miesiącach analogiczna współpraca została nawiązana z portalami DART-Europe (http://www.dart-europe.eu/) i ViFaOst (http://www.vifaost.de/)
- Wraz ze wzrostem zainteresowania wykorzystaniem danych z FBC przez zewnętrzne serwisy oraz tworzeniem nowych funkcji w FBC, brak spójności metadanych oraz konieczność upraszczania schematu przy kopiowaniu do bazy FBC zaczęły być coraz większym problemem

Metadane - krótko o historii…

- W drugiej połowie 2010 roku rozpoczęliśmy prace nad stworzeniem nowego schematu metadanych dla FBC
- Schemat ten miał z założenia być
  - łatwy do wykorzystania w komunikacji z większością polskich bibliotek cyfrowych
  - łatwy do wykorzystania w komunikacji z największymi serwisami zainteresowanymi wykorzystaniem danych z FBC
  - łatwy do wdrożenia w istniejących bibliotekach cyfrowych
    - względnie ☺ łatwy, gdyż proces migracji danych ze starego do nowego schematu może być również powiązany ze zmianą sposobu zapisu poszczególnych wartości, porządkowaniem słowników etc.
- W praktyce oznaczało to wykorzystanie elementów z uznanych schematów metadanych i rozszerzenie ich tam gdzie to potrzebne elementami specyficznymi dla polskich bibliotek cyfrowych

Opracowanie schematu metadanych dla FBC

- W naszej analizie wzięliśmy pod uwagę następujące schematy metadanych:
  - Dublin Core Metadata Element Set
    - http://www.dublincore.org/documents/dces/
  - Dublin Core Metadata Terms
    - http://dublincore.org/documents/dcmi-terms/
  - ETD-MS – profil DC do opisu prac dyplomowych i dysertacji
    - http://www.ndltd.org/standards/metadata/etd-ms-v1.1.html/
  - RIS – format obsługiwany przez wiele menedżerów bibliografii
    - http://www.refman.com/support/risformat_intro.asp
  - Highwire Press tags – jeden z formatów wspieranych przez Google Scholar
    - http://scholar.google.com/intl/en/scholar/inclusion.html

  oraz wyniki
  - wspomnianej wcześniej analizy schematów metadanych polskich bibliotek cyfrowych
  - dyskusji Zespołu ds. metadanych w bibliotekach cyfrowych

Schemat metadanych PLMET

- W efekcie powstał schemat metadanych nazwany **PLMET**
- Jest to schemat zalecany dla wszystkich bibliotek cyfrowych zainteresowanych współpracą z FBC
- Migracja FBC na schemat PLMET planowana jest na drugą połowę 2011 r.
- Jest to również preinstalowany schemat metadanych w dLibrze od wersji 5.0 tego systemu
- Dokumentacja schematu dostępna jest publicznie pod adresem: http://dl.psnc.pl/community/display/FBCMETGUIDE

Schemat metadanych PLMET

- PLMET docelowo ma stać się tzw. profilem aplikacji Dublin Core
- Specyfikacja takiego profilu określa:
  - Jaki jest cel/zastosowanie profilu
  - Jakiego typu obiekty opisuje ujęty w profilu schemat metadanych i jak są one między sobą powiązane
  - Jakie są elementy składowe schematu i na jakich zasadach powinny być używane
  - W jaki sposób metadane powinny być zapisane, aby można je przetwarzać w sposób zautomatyzowany

Por. http://dublincore.org/documents/profile-guidelines/

Schemat metadanych PLMET

- Schemat PLMET składa się z 59 elementów, z czego
  - 15 pochodzi ze schematu Dublin Core Metadata Element Set
    - Cały schemat DCMS
  - 33 pochodzą ze schematu Dublin Core Metadata Terms
    - Uszczegółowienia dla pól DCES: Tytuł (1), Zakres (2), Opis (2), Data (8), Format (2), Identyfikator (1), Powiązania (13), Prawa (2)
    - Pola głównego poziomu: Właściciel praw, Pochodzenie
    - Zrezygnowaliśmy z:
      - Pól specyficznych dla kolekcji dokumentów: accrualMethod, accrualPeriodicity, accrualPolicy
      - Pól związanych ściśle z kontekstem edukacyjnym: audience, educationLevel, instructionalMethod, mediator

Schemat metadanych PLMET

- Schemat ten składa się z 59 elementów, z czego
  - 5 pochodzi ze schematu ETD-MS – uszczegółowienie „Informacje o stopniu naukowym, zawodowym” dla pola Opis z DCES podzielone na 4 dodatkowe elementy:
    - Uzyskany tytuł
    - Stopień studiów
    - Dyscyplina
    - Instytucja nadająca tytuł
  - 6 uznaliśmy za specyficzne dla polskich bibliotek cyfrowych
    - Tagi użytkowników (uszczegółowienie pola Temat)
    - Miejsce wydania (uszczegółowienie pola Opis)
    - Sponsor digitalizacji (uszczegółowienie pola Opis)
    - Sygnatura (uszczegółowienie pola Identyfikator)
    - Digitalizacja (uszczegółowienie pola Pochodzenie)
    - Lokalizacja oryginału (uszczegółowienie pola Pochodzenie)

Schemat metadanych PLMET

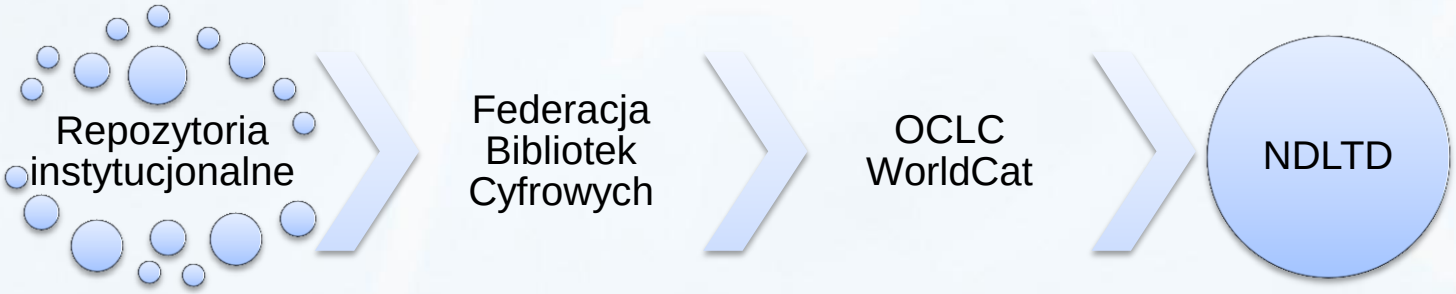
- Zrezygnowaliśmy z rozbijania pola „Cytata bibliograficzna” na podpola, co było istotne w kontekście wspomnianych schematów RIS czy Highwire i materiałów takich, jak artykuły z czasopism naukowych czy materiałów konferencyjnych
  - Chcieliśmy uniknąć w ten sposób rozszerzania schematu o kolejne kilka pól
  - Wybraliśmy podejście polegające na przyjęciu ustandaryzowanej formy zapisu cytaty bibliograficznej i automatycznej analizy takich zapisów w celu wydzielenia części składowych

Schemat metadanych PLMET

- 59 elementów zamiast 15 – czy to faktycznie potrzebne?
  - Istnienie 59 elementów **nie** oznacza konieczności ich wypełniania
  - Wprowadzenie uszczegółowień pól DCMS daje szansę na właściwe wykorzystanie doprecyzowanych semantycznie pól
    - Zakres – błędnie tłumaczony na „zasięg” i kojarzony np. z dostępnością publikacji, a nie z przestrzennymi lub czasowymi aspektami tematyki treści tej publikacji
    - Data – najczęściej wykorzystywana w znaczeniu daty wydania, ale nie zawsze…
    - Format – wykorzystywany do przechowywania informacji o formacie plików publikacji – pomieszanie metadanych opisowych i technicznych
    - Powiązania / Źródło – Brak spójnej koncepcji stosowania tych pól
    - Prawa – Bardzo często wykorzystywane wyłącznie do podania „właściciela” digitalizowanej publikacji

Schemat metadanych PLMET

- 59 elementów zamiast 15 – czy to faktycznie potrzebne?
  - Część uszczegółowień jest związana z pracami naukowymi
    - Wykorzystanie ETD-MS daje możliwość zbudowania następującego automatycznego przepływu metadanych



Schemat metadanych PLMET

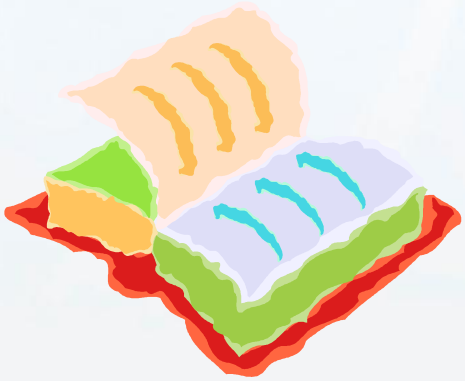
- 59 elementów zamiast 15 – czy to faktycznie potrzebne?
  - Część nowych uszczegółowień to pola już występujące w polskich bibliotekach cyfrowych, przydatne również na poziomie FBC
    - Tagi użytkowników (uszczegółowienie pola Temat)
    - Miejsce wydania (uszczegółowienie pola Opis)
    - Sponsor digitalizacji (uszczegółowienie pola Opis)
    - Sygnatura (uszczegółowienie pola Identyfikator)
    - Digitalizacja (uszczegółowienie pola Pochodzenie)
    - Lokalizacja oryginału (uszczegółowienie pola Pochodzenie)
  - Uszczegółowienia pól „Zakres” i „Powiązanie” oraz uszczegółowienie „Miejsce wydania” pola „Opis” dają szansę na łatwiejsze automatyczne wzbogacanie informacji przy agregowaniu ich w FBC

Schemat metadanych PLMET

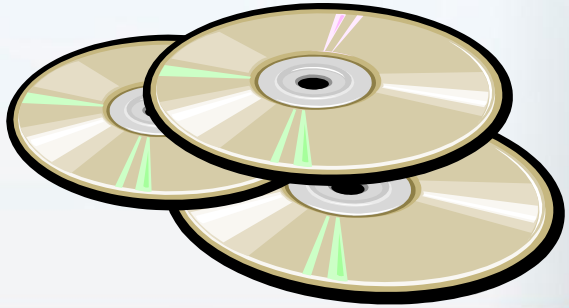
- Jeden z podstawowych problemów związanych z opisywaniem obiektów cyfrowych powstałych na skutek digitalizacji dotyczy łączenia opisu obiektu cyfrowego i obiektu źródłowego, który był digitalizowany
- W dokumencie „Using Dublin Core”

  http://www.dublincore.org/documents/usageguide/

  zaleca się stosowanie zasady jeden do jednego, czyli nie łączenie w jednym rekordzie opisu dwóch różnych obiektów
- Który obiekt w takim razie opisać – źródłowy czy cyfrowy?

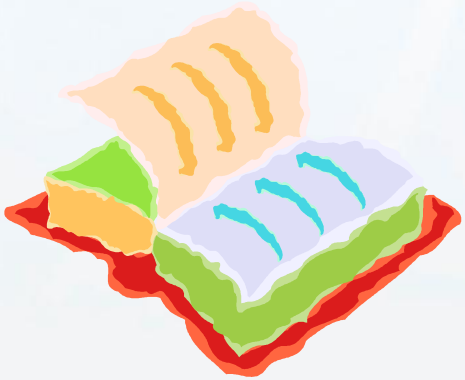


???

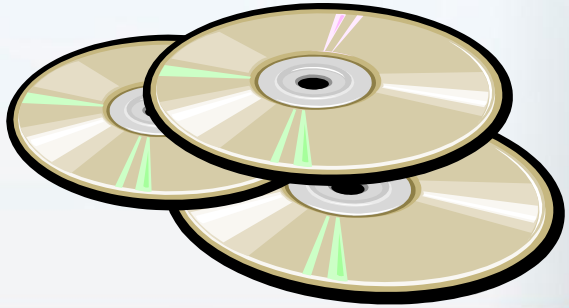


Schemat metadanych PLMET

- PLMET to schemat metadanych opisowych, które z natury dotyczą intelektualnej zawartości obiektu
  - Poza metadanymi opisowymi wyróżniamy jeszcze:
    - Metadane administracyjne (w tym techniczne)
    - Metadane strukturalne
- Przy podejmowaniu tego typu decyzji należy się kierować użytecznością biblioteki cyfrowej, a więc dobrem użytkowników końcowych
- Użytkownicy eksplorując zawartość biblioteki cyfrowej będą korzystali przede wszystkim z metadanych opisowych i wykorzystujących je funkcji biblioteki
- Który obiekt w takim razie opisać – źródłowy czy cyfrowy?



???



Schemat metadanych PLMET

Przykład na podstawie:
http://fbc.pionier.net.pl/id/oai:www.archiwumagnieszkiosieckiej.pl:45

Opis obiektu źródłowego:

- Tytuł: Dziennik ; [tom] XXVIII ; 17.04.1953 - 5.06.1953 r.
- Autor: Osiecka, Agnieszka (1936-1997)
- Typ zasobu: rękopisy
- Format: format A5
- Data (utworzenia): 17.04.1953 - 5.06.1953
- Identyfikator: [tom] XXVIII

Opis obiektu cyfrowego (zbioru plików):

- Tytuł: Dziennik ; [tom] XXVIII ; 17.04.1953 - 5.06.1953 r.
- Autor: Osiecka, Agnieszka (1936-1997)
- Typ zasobu: skan rękopisu
- Wydawca: Fundacja Okularnicy im. Agnieszki Osieckiej
- Format: image/x.djvu
- Data (wydania): 2009-07-07
- Identyfikator: oai:www.archiwumagnieszkiosieckiej.pl:45

Który z tych opisów jest bardziej przydatny dla użytkownika końcowego?

Schemat metadanych PLMET

- Który obiekt w takim razie opisać – źródłowy czy cyfrowy?
  - W ramach wytycznych do schematu PLMET, zakładając że metadane opisowe służą przede wszystkim odkrywaniu zawartości bibliotek cyfrowych, zalecamy aby dążyć do ujęcia w nich opisu obiektu źródłowego jako głównego obiektu
    - czyli np. podawać numer ISBN zeskanowanej książki w polu Identyfikator, a nie w polu Źródło czy Powiązanie
  - Informacje o obiekcie cyfrowym mają w tym kontekście charakter zdecydowanie bardziej administracyjny
    - Mogą być również opisane elementami schematu DC TERMS, jednak opis ten powinien być odrębny od metadanych opisowych

Schemat metadanych PLMET

- Skąd w takim razie elementy „Sponsor digitalizacji” czy „Digitalizacja” w schemacie PLMET?
  - Na chwilę obecną nie jesteśmy w stanie dokonać rozdziału metadanych opisowych i administracyjnych w pożądany sposób
  - Oprogramowanie wykorzystywane w bibliotekach cyfrowych (w tym również dLibra) nie posiada zazwyczaj możliwości elastycznego dostosowywania i udostępniania schematu metadanych administracyjnych
  - To wymusza konieczność przechowywania pewnych informacji administracyjnych w metadanych opisowych
  - W przyszłości pola takie jak „Sponsor digitalizacji” czy „Digitalizacja” powinny być przeniesione do dedykowanej sekcji administracyjnej
  - Podobnie może być z polami związanymi z informacjami prawnymi (o ile wiążą się one wyłącznie z obiektem cyfrowym) czy polem identyfikator w analogicznej sytuacji

Podsumowanie

- Mając na celu
  - ułatwienie dalszego rozwoju bibliotek cyfrowych w Polsce

  oraz
  - szeroką promocję gromadzonych w nich zasobów

  realizowane przez
  - automatyczne agregowanie metadanych z bibliotek cyfrowych
  - przetwarzanie tych metadanych, udostępnianie ich zewnętrznym serwisom i budowanie na ich podstawie nowych usług

  opracowaliśmy schemat metadanych o nazwie PLMET, z myślą o wdrożeniu go w najbliższych miesiącach w FBC oraz docelowo również w przyłączonych bibliotekach cyfrowych.

Podsumowanie

- W obecnej postaci PLMET jest schematem metadanych opisowych wraz z zestawem wytycznych
- Docelowo zostanie on najprawdopodobniej przekształcony w bardziej złożony schemat określający podział na pola związane z metadanymi opisowymi i administracyjnymi
- Podział ten ma na celu dokonanie wyraźnego rozróżnienia pomiędzy informacjami przydatnymi dla użytkownika przy odkrywaniu zasobów biblioteki cyfrowej (liczącej potencjalnie miliony obiektów) oraz informacjami przydatnymi na późniejszym etapie, gdy użytkownik pracuje już w kontekście znacznie mniejszej liczby obiektów

- Aktualna postać schematu PLMET jest dostępna pod adresem
  http://dl.psnc.pl/community/display/FBCMETGUIDE
- Zachęcam do zapoznania się z tą dokumentacją i nadsyłania komentarzy!

POZNAŃSKIE CENTRUM SUPERKOMPUTEROWO SIECIOWE

Dziękuję za uwagę - czy są pytania?

Kontakt ze mną:

Marcin Werla (mwerla@man.poznan.pl)

Poznańskie Centrum Superkomputerowo - Sieciowe
afiliowane przy Instytucie Chemii Bioorganicznej PAN,
ul. Noskowskiego 12/14, 61-704 Poznań,
tel : (+48 61) 858-20-00,
fax: (+48 61) 852-59-54,
e-mail: office@man.poznan.pl, http://www.man.poznan.pl