METADANE A PROCES CIĄGŁEJ DIGITALIZACJI

Agnieszka Wróbel
Biblioteka Uniwersytecka w Warszawie

Grzegorz Bednarek
GB Soft, sp.j. Zabrze

ALTO, biblioteka cyfrowa, DCMES, digitalizacja, EXIF, metadane deskryptywne, metadane administracyjno-techniczne, metadane strukturalne, METS, MIX, MODS, PREMIS, uniwersalne metadane wewnętrzne, XMP, biblioteka cyfrowa e-bUW

Podstawowym celem działalności, jaki przyświeca bibliotekom cyfrowym w Polsce, jest udostępnienie ich użytkownikom możliwie jak największej liczby materiałów w formie cyfrowej. Sukcesywna realizacja tego celu sprzyja upowszechnieniu się dostępu do dziedzictwa kulturowego oraz, co bardziej istotne, spełnia podstawowy wymóg, jaki postawiono bibliotekom cyfrowym, tj. ochronę dziedzictwa kulturowego.

Nie ulega wątpliwości, że w ostatnich latach – w procesach digitalizacji – znacząco wzrosła jakość wykonywanych metadanych i plików macierzystych. Zwiększa się zatem liczba instytucji, które, digitalizując swoje zbiory, przestrzegają coraz wyższych standardów skanowania z równoczesną dbałością o jednolitą postać metadanych. Jednocześnie, co może dziwić, kwestie związane z zarządzaniem, archiwizowaniem, ochroną czy też po prostu z zapoznawaniem się z wytworzonymi i gromadzonymi zasobami cyfrowymi podejmowane są bardzo rzadko. Wypada zatem zadać pytanie: czy w chwili obecnej działania takie podejmowane w większości bibliotek są wystarczające?

Metadane deskryptywne a proces digitalizacji obiektów bibliotecznych w e-bUW

Dostęp do zamieszczonych w bibliotece cyfrowej publikacji zależy przede wszystkim od jednolitych i poprawnych metadanych, ponieważ to ich zawartość gwarantuje efektywne i skuteczne wyszukiwanie zasobów. Miał to na uwadze zespół Biblioteki Uniwersyteckiej w Warszawie od momentu podjęcia decyzji o powołaniu do życia biblioteki cyfrowej e-bUW. W trakcie ustalania zasad or-

[1] Artykuł jest uzupełnioną wersją referatu wygłoszonego na międzynarodowej konferencji „Biblioteka cyfrowa dziś a wyzwania jutra", która odbyła się w dniach 24-25 stycznia 2013 r. w Bibliotece Jagiellońskiej w Krakowie.

ganizacyjnych przyjęto, iż standardy opisu dokumentów, wypracowane przez BUW podczas współkatalogowania w Narodowym Uniwersalnym Katalogu Centralnym NUKAT, będą stosowane również przez powstającą bibliotekę cyfrową. Koniecznym okazało się także określenie typu opisywanych dokumentów, a zatem wskazanie tego, czy opisy w bibliotece cyfrowej będą informacją bibliograficzną dotycząca dokumentów analogowych, czy też dokumentów zdigitalizowanych i zamieszczonych w e-bUW. Przyjęto, że w bibliotece cyfrowej opisywany będzie dokument analogowy, zaś udostępnione pliki publikacji cyfrowej traktowane będą jako kopia cyfrowa dokumentu analogowego. Jakkolwiek obecnie takie dylematy mogą zaskakiwać, kilka lat temu nie wszystkie projekty digitalizacyjne oparte były na powyższym założeniu.

Przyjęto, że w bibliotece cyfrowej e-bUW podstawą dla metadanych dokumentu cyfrowego będzie rekord dokumentu analogowego pochodzący z katalogu online bibliotek UW (OPAC), mimo iż ustalono, że podstawa analogowa każdego dokumentu cyfrowego zamieszczanego w bibliotece cyfrowej musi zostać uprzednio opracowana w katalogu NUKAT.

Jeżeli dla danego typu dokumentu nie zostały sporządzone odpowiednie instrukcje katalogowania, dokument opracowywany będzie w bazie lokalnej. Dodatkowym czynnikiem przemawiającym za tym, by metadane pobierać nie z katalogu NUKAT, lecz z katalogu OPAC jest fakt, że BUW rozpoczął katalogowanie swoich zbiorów w roku 1994[2], a co za tym idzie, pewna część opisów dostępna jest wyłącznie w katalogu lokalnym[3]. Ponadto zdecydowano, że te dokumenty, które zostaną wytypowane do zdigitalizowania, będą dodatkowo opisywane hasłami przedmiotowymi LCSH[4], które to wpisywane są do rekordów katalogu lokalnego, a nie katalogu centralnego.

Po opublikowaniu dokumentu cyfrowego w e-bUW rekord opisu bibliograficznego w katalogu NUKAT (a w przypadku dokumentu skatalogowanego lokalnie – w katalogu OPAC) uzupełniany jest o łącze do kopii cyfrowej. Pierwotnie udostępnione kopie cyfrowe dokumentów analogowych opisywane były w katalogu centralnym jako dokumenty elektroniczne. W konsekwencji ten sam dokument – w bazie danych – miał dwa opisy. Pierwszy jako dokument analogowy, zaś drugi jako dokument elektroniczny. Oba opisy zawierały w swych rekordach łącze do tej samej kopii cyfrowej. Okazało się, że takie rozwiązanie dla wielu czytelników było niezrozumiałe i budziło liczne nieporozumienia. W efekcie tego opisywanie opublikowanych dokumentów elektronicznych zarzucono. Obecnie opisywane są wyłącznie dokumenty analogowe, których rekord opisu bibliograficznego – po zdigitalizowaniu i udostępnieniu dokumentu elektronicznego – uzupełniany jest o łącze do wersji elektronicznej.

Biblioteka cyfrowa e-bUW – podobnie jak wiele polskich bibliotek cyfrowych – udostępnia zasoby cyfrowe za pośrednictwem systemu dLibra. Piąta

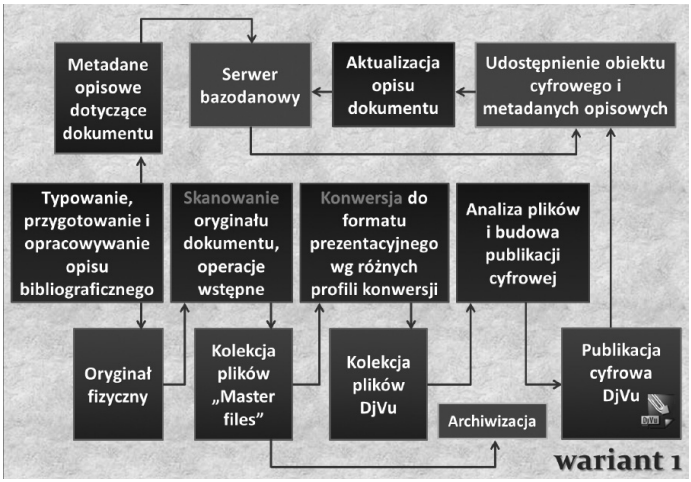
[2] NUKAT oficjalnie uruchomienie – 1 lipiec 2002 r. Pierwsze rekordy bibliograficzne w NUKAT były autorstwa BUW.

[3] Jednym z celów projektu „NUKAT – autostrada informacji cyfrowej" jest scalenie zasobów katalogów lokalnych z zasobami katalogu NUKAT. W ramach tego projektu na etapie automatycznego scalania, połączono informacje w 167 202 rekordach (stan na koniec II kwartału 2012 r.).

[4] Hasła przedmiotowe w języku angielskim nadawane są wszystkim dokumentom, które zlokalizowane są w Bibliotece Uniwersyteckiej w Warszawie w przestrzeni Wolnego Dostępu.

wersja dLibry pozwala opisywać dokumenty za pomocą standardu *Dublin Core* (DCMES)[5] lub w formacie PLMET[6].

W wersjach wcześniejszych dLibra umożliwiała tworzenie opisów wyłącznie w schemacie *Dublin Core*. W większości polskich bibliotek cyfrowych dla przyjętych schematów opisu publikacji podstawę stanowi właśnie standard *Dublin Core*, którego zestaw 15 atrybutów uzupełniany bywa o indywidualne rozwiązania (atrybuty) wynikające z określonych potrzeb bibliotek. Jako przykład przytoczyć można kwestie związane z wariantami tytułu dokumentu, opisem fizycznym dokumentu (pierwotnego), lokalizacją oryginału, itp. Obecnie biblioteka cyfrowa e-bUW udostępnia metadane oparte o standard *Dublin Core* rozszerzony o dodatkowe atrybuty: digitalizacja, sponsor digitalizacji, lokalizacja oryginału, opis fizyczny; i podatrybuty: miejsce wydania, wariant tytułu, identyfikator oryginału, słowa kluczowe, licencja, tekst licencji. Schemat tak zaplanowanej digitalizacji, przedstawiony został na rysunku 1.



Rys. 1. Schemat digitalizacji obiektów bibliotecznych w e-bUW

Cyfrowy produkt digitalizacji realizowanej w oparciu o zaprezentowany schemat to następujące pliki cyfrowe oraz metadane:

- pliki macierzyste (*master files*); w przypadku Biblioteki Uniwersyteckiej są to jednostronicowe pliki zapisane w formacie TIFF,
- pliki prezentacyjne (*digital publications*), a właściwie publikacje wielostronicowe zapisane w formacie DjVu i udostępniane za pośrednictwem Internetu,
- metadane deskryptywne przechowywane w bazie danych dLibry, a udostępniane w sposób dynamiczny, zależnie od potrzeb określonych przez osoby korzystające z zasobów biblioteki.

[5] Dublin Core Metadata Element Set (DCMES) w wersji 1.1 (simple DC).

[6] Federacja Bibliotek Cyfrowych. Dokumentacja schematu metadanych PLMET. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://dl.psnc.pl/community/display/FBCMETGUIDE>.

Korzyści z posiadania metadanych opisowych dotyczą wyłącznie możliwości przeszukiwania zasobów biblioteki, lokalizowania poszukiwanej publikacji lub też wyświetlania listy publikacji spełniających określony przez czytelnika warunek.

Metadane wewnętrzne w publikacjach cyfrowych w e-bUW

W marcu 2009 r., w ramach „Programu Operacyjnego Innowacyjna Gospodarka, lata 2007-2013", Biblioteka otrzymała pozytywną decyzję o finansowaniu projektu „NUKAT – autostrada informacji cyfrowej". Projekt ten składał się z czterech zadań:

- rozbudowa centralnej informacji w NUKAT poprzez scalanie danych z katalogów lokalnych, wykonanych przed uruchomieniem katalogu NUKAT,
- modernizacja i rozwój wyszukiwarki KaRo pozwalającej na przeszukiwanie katalogów bibliotek nieobjętych współpracą z NUKAT-em,
- retrokonwersja 1600 tytułów czasopism znajdujących się w zbiorach BUW,
- cyfryzacja najcenniejszych czasopism ze zbiorów BUW i umieszczenie ich w bibliotece cyfrowej Uniwersytetu Warszawskiego e-bUW.

Ostatnie zadanie projektu to zeskanowanie 3 000 000 stron (do plików w formacie TIFF) czasopism pochodzących głównie z XIX i XX w., powstałych w znacznej części na kwaśnym papierze, a następnie przetworzenie tak powstałych plików TIFF do wielostronicowych publikacji w prezentacyjnym formacie DjVu. Ponieważ format plików DjVu, podobnie jak i format pdf, są formatami hybrydowymi, przyjęto, że wszystkie publikacje DjVu wyposażone będą w ukryte warstwy rozpoznanego tekstu uwzględniające logiczną strukturę treści kolejnych stron publikacji. W publikacjach cyfrowych tekst nie został zapisany według przyjętej przez użyty do rozpoznania OCR program kolejności poszczególnych części stron czy kolumn, tekst zapisano tak, by treści kolejnych artykułów zawartych na pojedynczej stronie publikacji były ciągłymi i zwartymi blokami, zapisanymi w ukrytej warstwie tekstowej kolejno po sobie. Ponadto przyjęto, że do wszystkich publikacji, które mają spisy treści, dodane będą wizualne środki nawigacji po dokumencie, nazywane mapowaniem publikacji.

W chwili rozpoczęcia prac nad projektem „NUKAT – autostrady informacji cyfrowej", metadane deskryptywne były jedyną kategorią metadanych, jakim poświęcono uwagę i jakie zamierzano gromadzić wraz z cyfrowymi postaciami zdigitalizowanych obiektów. Przyjęto zasadę, że podczas ciągłej digitalizacji obiektów bibliotecznych moment, w którym oryginał zostaje przekazany do skanowania, jest jednocześnie momentem, w którym metadane opisowe dostępne będą w systemie dLibra jako rekord publikacji planowanej.

Szybko narastająca ilość plików cyfrowych była przyczyną podjęcia działań, które pozwalały zweryfikować kompletność przekazywanych partii plików, jakości ich wykonania, efektywnego zarządzania plikami macierzystymi oraz plikami prezentacyjnymi, sprawdzania dostępności przygotowanych w Bibliotece Uniwersyteckiej metadanych deskryptywnych (dla zdigitalizowanych już obiektów), wykonywania kopii bezpieczeństwa plików, itp. Jakkolwiek działania te realizowane były pomyślnie, odczuwalny był brak narzędzi i środków pozwalających w zadowalający sposób dokumentować ich wyniki. Dzięki sukcesywnie

tworzonym metadanym deskryptywnym satysfakcjonujący był natomiast stan opisów zdigitalizowanych obiektów, zaistniała jednak potrzeba opracowania sposobu przechowywania kolejnych (obok metadanych deskryptywnych) informacji na temat zdigitalizowanych obiektów. Niezbędne są zatem metadane dotyczące archiwizowanych plików oraz udostępnianych publikacji cyfrowych, zasad i praw do korzystania z udostępnianych zasobów cyfrowych.

Pod koniec roku 2010 r. jeden z wykonawców digitalizacji, spółka GB Soft[7], złożyła propozycję[8] dotyczącą sposobu przechowywania metadanych nie w posiadanych lub zupełnie nowych bazach danych, lecz bezpośrednio w plikach publikacji cyfrowych. Tym samym możliwym stało się gromadzenie i przechowywanie zupełnie nowych kategorii metadanych, które do tej pory nie były stosowane w Bibliotece Uniwersyteckiej w Warszawie. Istotną cechą tej propozycji było to, iż nie wymagało to zmiany w BUW infrastruktury informatycznej, a zatem poniesienia nakładów finansowych, nie powodowało też opóźnienia całego procesu digitalizacyjnego.

Początkowo propozycja dotyczyła umieszczenia w pliku publikacji cyfrowej wyłącznie kompletu metadanych deskryptywnych odpowiadających takim atrybutom, które czytelnikowi udostępnia także dLibra (np. pod postacią tekstowego pliku RDF), później została rozszerzona o możliwość zapisu w plikach publikacji cyfrowych metadanych technicznych oraz metadanych administracyjnych dotyczących proweniencji plików DjVu, które to względem plików macierzystych (w formacie TIFF) traktowane są jako pliki pochodne. Umieszczone w publikacjach DjVu metadane techniczne opisywały kolejno każdą stronę dokumentu i to zarówno w odniesieniu do pliku w formacie macierzystym, jak i w formacie prezentacyjnym. Z kolei metadane administracyjne informowały o zastosowanym oprogramowaniu oraz sposobie przetworzenia (konwersji) strony w formacie macierzystym do strony w formacie prezentacyjnym. Rejestrowane były ponadto daty kolejnych czynności wykonywanych podczas digitalizacji, tj. daty skanowania, konwersji oraz dodania metadanych do przygotowanych publikacji a także nazwa Biblioteki Uniwersyteckiej jako właściciela praw do plików macierzystych.

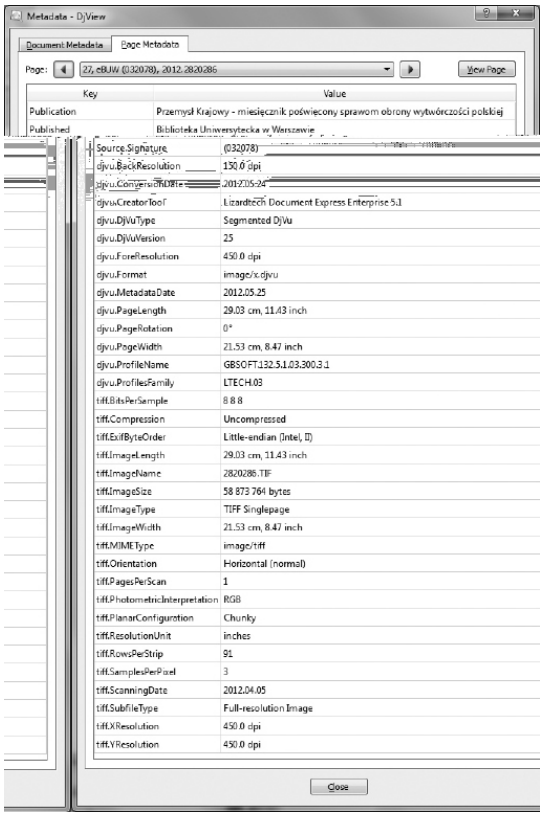
Propozycja wykonawcy została zaakceptowana przez BUW, w efekcie w metadane deskryptywne wyposażone zostało około 2 750 000 plików stron w formacie prezentacyjnym, wśród których około 2 200 000 wyposażono również w metadane administracyjno-techniczne.

Rozmiar publikacji cyfrowych w formacie DjVu, po wyposażeniu ich w metadane opisane powyżej, zwiększył się nieznacznie, dla przeciętnego czasopisma 8-16-stronicowego o około 10 kB. Udostępnianie publikacji cyfrowych wyposażonych w zintegrowane metadane wewnętrzne (czyli umieszczone w plikach publikacji) pozwala czytelnikom pobrać nie tylko samą publikację, ale jednocześnie również i wszystkie dokumentujące ją metadane, w tym metadane deskryptywne. Publikacje posiadające metadane wewnętrzne – podczas

[7] GB Soft, Zabrze. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.djvu.com.pl>.

[8] Propozycja nie była elementem przedmiotu zamówienia określonego zarówno w dokumencie SIWZ jak i podpisanej umowie dotyczącej sposobu wykonania publikacji cyfrowych.

zapoznawania się z ich zawartością, czy też podczas ich wydruku całkowitego bądź częściowego – niczym nie różnią się od publikacji, które takich metadanych nie posiadają, natomiast dzięki narzędziom informatycznym pozwalają np. wyświetlić komplet zintegrowanych w nich metadanych. Do narzędzi takich należy np. przeglądarka plików DjVu – *DjView*, która jest składową pakietu *DjVu Libre*[9] lub przeglądarka metadanych wewnętrznych (dla bardzo dużej liczby formatów plików) *ExifTool*[10].



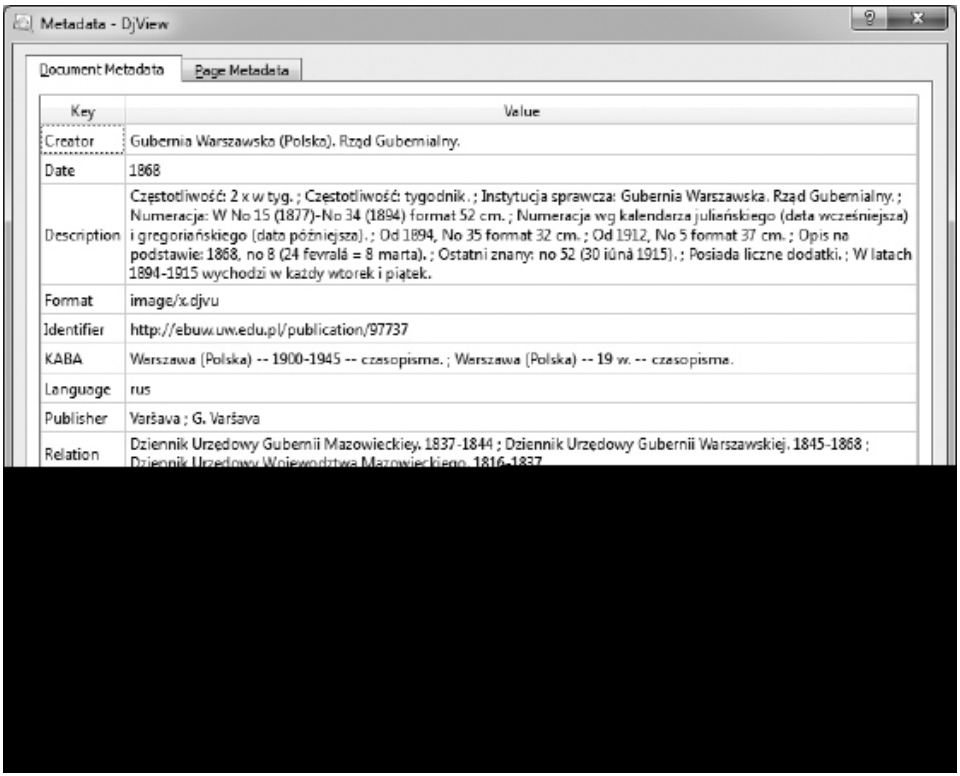
Rys. 2. Metadane techniczne zintegrowane w publikacji cyfrowej

Dostęp do metadanych wewnętrznych jest prosty, korzystanie z nich nie nastręcza żadnych kłopotów, a ich utworzenie nie wymaga dużego nakładu pracy. Stanowią też wygodną formę do przechowywania w zasobach cyfrowych biblioteki. Wewnętrzne metadane mogą posłużyć jako narzędzie pozwalające weryfikować wiarygodność i rzetelność wykonanych publikacji cyfrowych oraz plików macierzystych (względem oryginału papierowego). W prosty sposób mogą zostać wyeksportowane z publikacji cyfrowych do plików w określonym formacie i być dalej przetwarzane, na przykład w celu weryfikacji kompletności

[9] Sourceforge. DjVuLibre [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://sourceforge.net/projects/djvu/>.

[10] ExifTool by Phil Harvey. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.sno.phy.queensu.ca/~phil/exiftool/>.

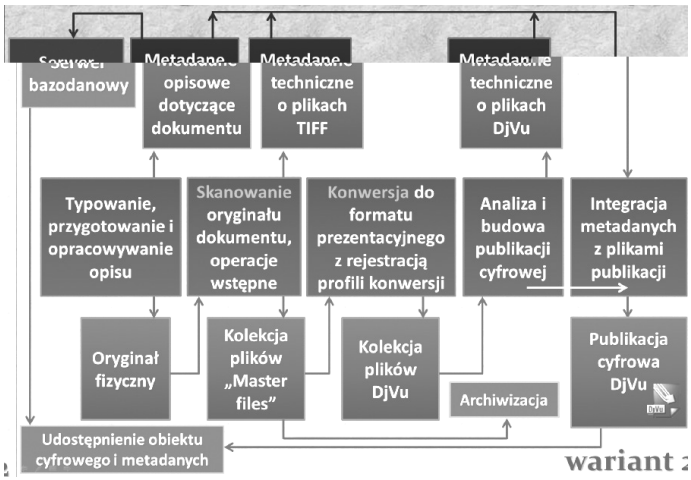
archiwizowanych kolekcji plików macierzystych. To, czy formatem określonym dla eksportu wewnętrznych metadanych będzie format plików .csv[11], prosty plik tekstowy, czy też inny format, zależy wyłącznie od osoby wykonującej tę czynność. Dzięki dostępności informacji o sposobie konwersji poszczególnych plików macierzystych do formatu prezentacyjnego metadane wewnętrzne są również bazą wiedzy o tym, w jaki sposób prowadzić w przyszłości kolejne digitalizacje, by osiągnąć założony poziom jakości publikacji w formacie prezentacyjnym. Przykładowy wygląd okienka wyświetlającego wewnętrzne metadane techniczne pojedynczej strony publikacji oraz metadane deskryptywne publikacji cyfrowej zaprezentowano na rysunkach 2 i 3.



Rys. 3. Metadane deskryptywne zintegrowane w publikacji cyfrowej

Przed wprowadzeniem takiego sposobu dokumentowania obiektów cyfrowych Biblioteka Uniwersytecka gromadziła jedynie metadane deskryptywne, które przechowywane były wyłącznie w bazach danych. Pomimo istotnej zmiany w sposobie ciągłej produkcji obiektów cyfrowych schemat digitalizacji, który przedstawiono na rys. 4, a uwzględniający gromadzenie metadanych administracyjno-technicznych oraz integrowanie ich w plikach publikacji cyfrowych, nie uległ znaczącym zmianom w stosunku do schematu digitalizacji zaprezentowanego na rys. 1.

[11] Format plików comma separated values.



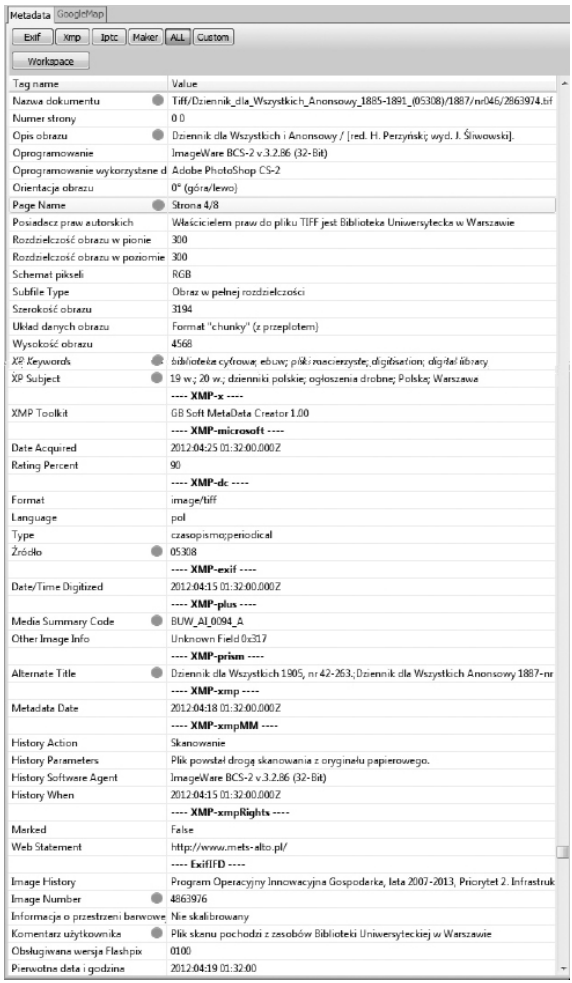
Rys. 4. Schemat digitalizacji obiektów bibliotecznych uwzględniający integrowanie metadanych wewnętrznych

Specyfikacja formatu plików DjVu pozwala integrować w plikach publikacji metadane o całej publikacji (cechę tę wykorzystano dla metadanych deskryptywnych) oraz metadane o każdej stronie publikacji z osobna (zastosowano dla metadanych administracyjno-technicznych). Metadane te mogą być integrowane w publikacjach DjVu dwoma sposobami: albo pod postacią kontenera metadanych „*DjVu meta*" albo pod postacią kontenera metadanych zgodnego ze standardem firmy Adobe – XMP[12] (kontener metadanych to komplet zapisanych zgodnie z wymogami określonymi w specyfikacji formatu plików DjVu).

W ramach projektu „NUKAT – autostrada informacji cyfrowej", publikacje DjVu, w których zintegrowano metadane, posiadają metadane wewnętrzne zapisane zarówno pod postacią kontenerów „*DjVu meta*" jak i kontenerów XMP, co gwarantuje możliwie największą ich elastyczność oraz funkcjonalność. W momencie digitalizacji nie można było przewidzieć, jakie powstaną w przyszłości aplikacje dla formatu DjVu i z której postaci wewnętrznych metadanych będą potrafiły skorzystać. Przykładowo, niektóre opcje systemu operacyjnego Windows7 korzystają z wewnętrznych metadanych zapisanych pod postacią kontenerów „*DjVu meta*" i jednocześnie nie potrafią skorzystać z metadanych wewnętrznych zapisanych jako kontenery XMP.

Przykładowy kontener XMP metadanych wewnętrznych, zintegrowany w pliku macierzystym (jednej ze stron publikacji cyfrowej), wyświetlony za pomocą przeglądarki *ExifTool*, przedstawiono na rys. 5.

[12] Extensible Metadata Platform (XMP). [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.adobe.com/products/xmp/>.



Rys. 5. Wewnętrzne metadane zapisane jako kontener XMP

Metadane administracyjno-techniczne, w które została wyposażona znacząca ilość stron wykonanych publikacji cyfrowych (jak już wspomniano – około 2 200 000 stron, co stanowi ponad 70% zdigitalizowanego zasobu), mają taką samą postać jak metadane deskryptywne, tzn. są kolekcją par: atrybut – wartość atrybutu. Wartością danego atrybutu może być opis, wartość numeryczna, data, a niektórym spośród z nich – w określonym przypadku – można nie przypisywać żadnej wartości. Podobnie jak w przypadku metadanych deskryptywnych, niektórym atrybutom przypisać można wyłącznie pojedynczą wartość, innym zaś wiele wartości. Lista wartości atrybutów – również jak w przypadku metadanych deskryptywnych – dla pewnych atrybutów może być listą uporządkowaną, nieuporządkowaną bądź listą wartości, dla których dodatkowo określono języki, w których wprowadzono kolejne pozycje takiej listy. Taka postać metadanych odpowiada wprost opisowi metadanych wykonanych dla przykładu jako tekstowe pliki xml odpowiadające specyfikacji RDF.

Zarówno przeglądarka *DjView* (rys. 2 oraz rys. 3), jak i przeglądarka *Exiftool* (rys. 4) w taki właśnie wizualny sposób pozwalają zapoznawać się z wewnętrznymi metadanymi analizowanej publikacji.

Podstawą wykonania metadanych deskryptywnych w e-bUW jest standard metadanych *Dublin Core*, natomiast dla wewnętrznych metadanych administracyjno-technicznych podstawę taką stanowią dwa dominujące dla plików obrazów standardy metadanych – EXIF[13] oraz wspomniany już standard XMP. Dzięki rozszerzeniu procesu gromadzenia metadanych – podczas digitalizacji – publikacje cyfrowe e-bUW wyposażone zostały w:

- **Metadane deskryptywne**, powstałe w oparciu o standard *Dublin Core,*
- **Metadane administracyjno-techniczne**, powstałe w oparciu o standardy EXIF oraz XMP, wśród których ze względu na zawartość wydzielić można następujące kategorie metadanych:
  - dotyczące praw autorskich oraz zasad udostępniania publikacji cyfrowych,
  - dotyczące sposobu ekspozycji oraz użytkowania publikacji cyfrowych,
  - techniczne,
  - dotyczące proweniencji i wzajemnych relacji pomiędzy poszczególnymi plikami cyfrowymi dotyczącymi tego samego obiektu,
  - związane z procesem archiwizacji plików cyfrowych.

Metadane METS

W roku 2011[14] w ramach Wieloletniego Programu Rządowego „Kultura+" Biblioteka Narodowa, jako Centrum Kompetencji w zakresie digitalizacji materiałów bibliotecznych, wprowadziła nowe warunki przekazywania jej obiektów cyfrowych, zdigitalizowanych w ramach tego programu. Jednym z warunków był wymóg wykonania dla przekazywanych do BN obiektów cyfrowych metadanych deskryptywnych, technicznych, administracyjnych oraz metadanych dotyczących struktury fizycznej i logicznej zdigitalizowanych obiektów.

W celu gromadzenia tak wielu i tak dalece różniących się informacji na świecie oferowanych jest kilka standardów metadanych. Z punktu widzenia potrzeb instytucji zajmujących się digitalizacją uwagę skupić można na standardach zaproponowanych przez Bibliotekę Kongresu Stanów Zjednoczonych (LOC)[15]. Standardy te – w procesie przygotowywania kompletu metadanych o danym obiekcie – uzupełniają się, choć dla niektórych właściwości obiektu stosowane mogą być niekiedy zamiennie.

Standardami tymi są:

- METS – Metadata Encoding and Transmission Standard,
- MODS – Metadata Object Description Schema,
- MIX – NISO Technical Metadata for Digital Still Images,
- PREMIS – The PREMIS Data Dictionary for Preservation Metadata,
- ALTO – Analyzed Layout and Text Object.

[13] EXIF.org. Exchangeable Image File Format. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.exif.org/>.

[14] Warunki doprecyzowano w 2012 r.

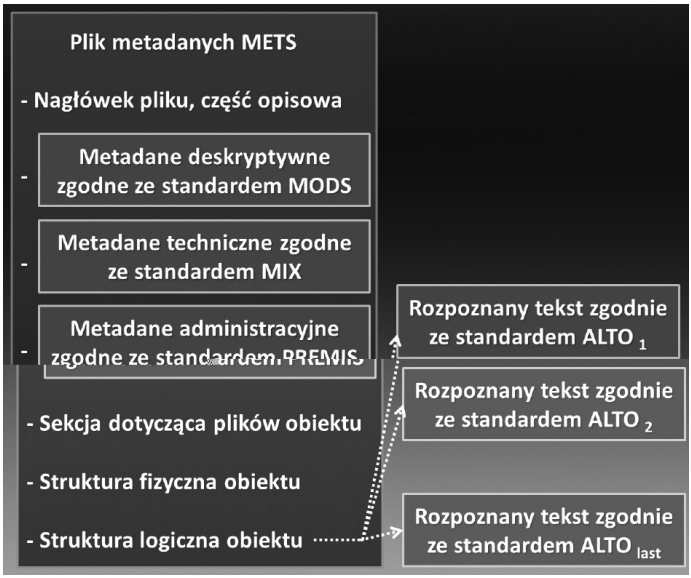
[15] Library of Congress [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.loc.gov/index.html>.

Najważniejszym standardem metadanych jest standard METS. Być może właśnie dlatego kompletnie opracowane metadane obiektu bywają nazywane „metadanymi METS". Metadane opracowuje się pod postacią plików tekstowych zapisanych zgodnie z wymogami języka xml i kodowanych jako pliki UTF–8. Metadane opracowane w oparciu o standardy METS, MODS, MIX i PREMIS przechowywane są zazwyczaj w pojedynczym (zbiorczym) pliku tekstowym. Natomiast rozpoznany na drodze operacji OCR tekst kolejnych stron publikacji przechowywany jest najczęściej w kolejnych plikach tekstowych odpowiadających wymogom standardu ALTO.

Tekstowe pliki metadanych METS (wraz z towarzyszącymi im plikami metadanych ALTO) mogą służyć zachowywaniu zawartych w nich metadanych, jednak najczęściej za pośrednictwem odpowiedniego systemu informatycznego - po zakończeniu procesu digitalizacji – zawartość tych plików importowana jest do bazy danych przechowującej wszystkie kategorie metadanych o wszystkich zdigitalizowanych dokumentach.

Wymóg Biblioteki Narodowej w ramach Programu „Kultura+" zakładał wykorzystanie powyższych standardów.

Plik metadanych METS składa się z kilku części następujących jedna po drugiej i niemieszających się wzajemnie w jakimkolwiek przypadku. Budowę pliku metadanych METS przedstawia rys. 6.



Rys. 6. Budowa pliku metadanych METS

● **Nagłówek pliku METS** zawiera kilka ogólnych informacji na temat repozytorium, obiektu, jego właściciela oraz wszystkie użyte w pliku METS przestrzenie nazw.

● (autonomiczna część pliku) **metadanych deskryptywnych** zawiera metadane deskryptywne dotyczące opisu całego obiektu, może też zawierać opis całego obiektu oraz opisy jego poszczególnych części (np. po-

szczególnych artykułów, zamieszczonych zdjęć lub ilustracji). Zalecane jest, by takie metadane opracowywać zgodnie ze standardem MODS. Możliwym do zastosowania jest również standard DCMES albo MARC XML. Dopuszcza się również opracowanie metadanych deskryptywnych jednocześnie w więcej niż jednym standardzie, np. wg standardu MODS oraz DCMES. Ponieważ niektóre atrybuty standardu DCMES nie są dostępne wprost w standardzie MODS (np. atrybut *relation*), opcja ta może okazać się bardzo przydatna. Jeżeli dla danego obiektu dostępny jest kompletny opis odpowiadający standardowi DCMES, może on być użyty jako standard, którego metadane uzupełnią opis obiektu wykonany za pomocą metadanych zapisanych w standardzie MODS.

● **Kontener metadanych technicznych** przygotowywany jest najczęściej w oparciu o standard metadanych MIX. Dla metadanych technicznych jest to standard zalecany. Ponieważ możliwości standardu PREMIS dotyczą również niektórych właściwości plików cyfrowych, zdarza się, że niekiedy standard PREMIS używany jest jako zamiennik standardu MIX. Wybór standardu zależy od opisywanych metadanych, a decyduje o nim osoba opracowująca metadane techniczne.

● **Kontener metadanych administracyjnych (konserwatorskich)** zawiera przede wszystkim informacje o plikach pochodnych względem plików macierzystych, a właściwie nie tyle o plikach, co o sposobie i przyczynie ich powstania oraz o ewentualnych ich modyfikacjach. Plikami pochodnymi są np. pliki udostępnianych publikacji cyfrowych oraz pliki przechowujące rozpoznany na kolejnych stronach publikacji tekst. Ponadto kontener ten powinien zawierać również wszelkie metadane związane z opisem możliwego sposobu wykorzystania udostępnionej publikacji oraz praw autorskich związanych z opisywaną publikacją cyfrową. Metadane administracyjne opracowywane są zgodnie ze standardem PREMIS.

● W niektórych przypadkach w pliku METS może wystąpić kontener metadanych niewyszczególniony na rys. 6. Dotyczy to przede wszystkim obiektów, które posiadają znikome lub nie posiadają w ogóle jakichkolwiek ograniczeń prawnych co do sposobu ich wykorzystywania. Wtedy wygodne jest opracowanie informacji dotyczących praw do obiektu w osobnym kontenerze metadanych – często nazywanym RIGHTS, zgodnie z rozwiązaniem zaproponowanym przez *Stanford University Libraries*[16]. Rozwiązanie to jest szczególnie wygodne w przypadku polskich bibliotek cyfrowych, w których zdecydowana większość obiektów cyfrowych dotyczy domeny publicznej.

● **Szczegółowe zestawienie wszystkich plików** (nie tylko macierzystych) zdigitalizowanego obiektu papierowego. W tej sekcji pliku METS umieszczone są nazwy plików, daty ich utworzenia, rozmiary, sumy kontrolne, lokalizacje i kilka innych informacji. Sekcja ta najczęściej opisuje trzy do pięciu grup plików:

● *Master files* – pliki obiektu w formacie macierzystym (np. w formacie TIFF),

● *Reference copy* – pliki w formacie prezentacyjnym, gdy do udostępniania publikacji wybrany został format plików jednostronicowych (np. JPEG, JPG2000, PNG),

● *Digital publications* – pliki publikacji wielostronicowych w formacie prezentacyjnym (np. DjVu),

[16] Stanford University Libraries [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web:<http://cosimo.stanford.edu/sdr/>.

● *Recognized Text* – tekstowe pliki xml zapisane zgodnie ze standardem ALTO, a zawierające m.in. rozpoznany w plikach macierzystych tekst,

● *Thumbnails* – pliki miniatur graficznych kolejnych stron publikacji (np. GIF, JPG).

● **Sekcja** informacji dotyczących **struktury fizycznej obiektu**. Kolejna sekcja pliku METS, w której opisane są relacje pomiędzy poszczególnymi plikami obiektu (np. porządek, w jakim powinny być udostępniane).

● **Sekcja** informacji dotyczących **struktury logicznej publikacji cyfrowej**. Ostatnia sekcja w pliku METS zawiera informacje związane z zawartością obiektu cyfrowego. Przykładowo, dla plików audio mogą być opracowane metadane dotyczące podziału audycji na kolejne jej części lub tematy, dla plików wideo dostępne mogą być informacje o kolejnych scenach pliku, zaś dla plików zagitalizowanej książki lub czasopisma mogą to być informacje dotyczące poszczególnych rozdziałów, artykułów, rozmieszczenia ich tytułów, użytych czcionek, występujących w publikacji zdjęć, tabel lub ilustracji.

Jeżeli podczas digitalizacji wykonywane były operacje rozpoznania tekstu zdigitalizowanych stron oryginału obiektu (OCR), plikowi metadanych METS dotyczącemu tego obiektu towarzyszą pliki tekstowe odpowiadające wymogom standardu ALTO, w których zachowywany jest m.in. rozpoznany tekst. Rozpoznany tekst względem obiektu, którego dotyczy, pełni rolę metadanych.

Wykonanie plików metadanych METS jest procesem wymagającym znaczącego nakładu pracy. Nakład pracy niezbędny dla wykonania metadanych MODS, MIX, PREMIS czy też metadanych dotyczących struktury fizycznej jest dla większości obiektów podobny. Najbardziej pracochłonne jest wykonanie metadanych dotyczących struktury logicznej obiektu oraz metadanych ALTO. Pliki ALTO zawierają nie tylko rozpoznane na stronie słowa wraz z kompletami współrzędnych pozwalającymi wskazać położenie tych słów na stronie, ale również informacje powstałe w wyniku analizy graficznej zawartości stron. Poza podstawowymi informacjami zachowywanymi w plikach ALTO, do których zalicza się obszary zajęte przez tekst oraz przez marginesy, rodzaje, wielkości i kolory użytych na stronie rodzin czcionek drukarskich, w plikach ALTO zawarte są ponadto informacje o:

● poszczególnych częściach obiektu (rozdziały, artykuły, wstawki reklamowe, itp.),

● układzie tekstu (szpalty, nagłówki, stopki redakcyjne, zestawienia tabelaryczne, itp.),

● elementach graficznych (ryciny, drzeworyty, zdjęcia, wykresy, itp.).

Każdy współtworzący zawartość strony element, dla przykładu artykuł, opisany może być albo tytułem, jaki mu nadano oraz treścią samego artykułu, albo za pomocą fragmentów, z których się składa i z podtytułów, jakie tym fragmentom nadano. Taki rekurencyjny sposób opisu rozdziałów czy też artykułów zawartych w obiekcie cyfrowym jest identyczny ze sposobem, w jaki opisuje się foldery. Foldery posiadają swój tytuł (nazwę), zawartość oraz mogą posiadać (pod) foldery, które też mają tytuł, zawartość oraz mogą mieć własne (pod)foldery itd.

Poziom szczegółowości, z jaką należy wykonać analizę zawartości poszczególnych stron obiektów jest autonomiczną decyzją instytucji planującej sposób, w jaki przeprowadzona będzie digitalizacja zasobów. Dziesięciokrotnie

większy nakład pracy bardzo szczegółowej analizy zawartości stron względem podstawowej – jedno- lub co najwyżej dwupoziomowej – analizy tych stron nie będzie w żadnym wypadku nieoczekiwanym wzrostem koniecznego nakładu pracy. Niektóre biblioteki europejskie kilka spośród elementów współtworzących strony (np. czasopism) uznały za elementy podstawowe, dla których wykonanie analizy logicznej jest obligatoryjne i wykonane musi być w oparciu o przyjęte procedury standardowe. Przykładem może być Biblioteka Narodowa Księstwa Luksemburg[17], gdzie w postaci cyfrowej produktu digitalizacji takie elementy stron czasopism jak prognozy pogody, nekrologi oraz reklamy są różnymi elementami stron – z punktu widzenia sposobu wykonania struktury logicznej – i wykonywane muszą być według ściśle określonych procedur.

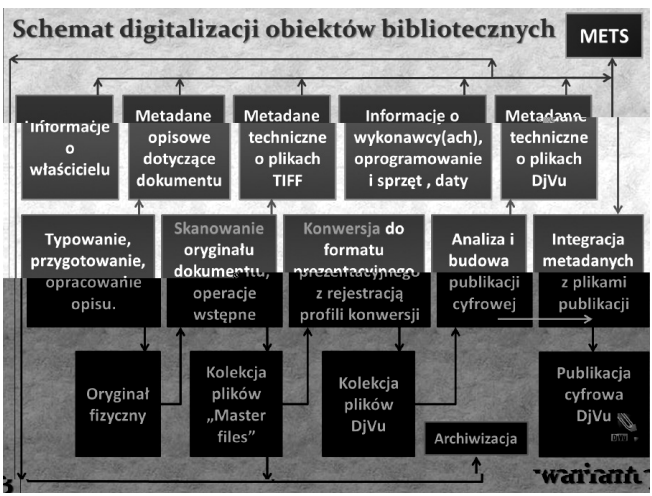
Rozmiar powstałego pliku METS (bez towarzyszących mu plików ALTO) dla kilkunastostronicowego czasopisma może przekroczyć 100 kB. Dla obiektów 200- lub 300-stronicowych rozmiar pliku METS przekraczający 5 MB nie jest niczym nadzwyczajnym. Dla porównania można wskazać, że metadane *Dublin Core* opisujące dowolny obiekt biblioteczny, pobrane z witryny dowolnej biblioteki cyfrowej jako plik RDF, mają rozmiar 2-7 kB, dlatego też podjęcie decyzji o tym, w jaki sposób należy prowadzić opracowywanie metadanych dla digitalizowanych obiektów, oparte musi być na szczegółowej analizie ich przydatności oraz wiedzy o tym, czy przyjęty scenariusz opracowywania metadanych został również wprowadzony w innych instytucjach, a przede wszystkim, czy spełnił oczekiwania. Innymi słowy, że realizacja potrzeby dokumentowania obiektów cyfrowych za pomocą różnych kategorii metadanych to konieczność, ale należy przyjąć rozwiązanie odpowiadające przyjętym na świecie standardom, bo w przeciwnym razie z zasobów danych nie będą mogły skorzystać np. serwisy internetowe – zewnętrzne względem biblioteki, która takie metadane wykonała – i zasięg ich wykorzystania będzie ograniczony.

Ponieważ obecnie w Polsce program „Kultura+” jest albo jedynym, albo jednym z nielicznych programów, w ramach którego postawiono wymóg opracowywania metadanych innych niż tylko deskryptywne, w celu porównania zapoznano się z preferencjami, według których opracowywane są metadane w wybranych biblioted innm lb nUciw hwh -5

wości – wykonania metadanych (aż 50% wagi w ocenie ofert za doświadczenie zawodowe oferenta, zrealizowane digitalizacje, które odpowiadały wszystkim wymogom ocenianej oferty).

Decyzja, by digitalizację prowadzić tak, by posiadać kompletne metadane odpowiadające standardom METS / MODS, DCMES / MIX / PREMIS / ALTO, jest – wbrew pozorom – decyzją, która jednocześnie upraszcza i porządkuje wiele czynności składających się na proces ciągłej digitalizacji oryginałów obiektów. Planując wykonanie pliku metadanych METS, należy z góry zaplanować odpowiednio jednolitą, spójną i czytelną strategię nazewnictwa dla tysięcy – różniących się formatem oraz zawartością – plików cyfrowych, które powstaną w trakcie digitalizacji oraz nazewnictwa folderów, w których będą one przechowywane. Jeżeli w scenariuszu digitalizacji zgromadzono odpowiednią ilość wymaganych informacji (pod postacią wygodną dla wykonawcy digitalizacji), sam proces ciągłego generowania tekstowych plików metadanych METS może być wykonany w innym momencie i nie zachodzi konieczność tworzenia pojedynczych plików METS w ślad za sukcesywnie digitalizowanymi pojedynczymi obiektami cyfrowymi.

Schemat digitalizacji w Bibliotece Uniwersyteckiej w Warszawie, w ramach projektu „NUKAT – autostrada informacji cyfrowej" został ponownie zmodyfikowany, tak by umożliwiał nie tylko gromadzenie metadanych deskryptywnych, administracyjno-technicznych oraz strukturalnych, ale by możliwe było – w razie potrzeby – opracowanie i wykonanie na ich podstawie plików metadanych METS. Scenariusz digitalizacji przedstawiono na rys. 7.



Rys. 7. Schemat digitalizacji obiektów bibliotecznych umożliwiający wykonywanie plików metadanych METS

Repozytoria cyfrowe, w których zgromadzone kolekcje plików udokumentowano za pomocą metadanych METS, mają – co oczywiste – znacznie więcej możliwości operowania, wykorzystywania bądź zarządzania własnymi zasobami niż repozytoria, których obiekty cyfrowe opisano wyłącznie metadanymi deskryptywnymi.

Przykładowo: realizowana w brytyjskich archiwach TNA (*The National Archives*)[18] digitalizacja kilkunastu milionów stron obiektów prowadzona jest tak, że w zależności od digitalizowanego obiektu produkt cyfrowy może mieć jedną z dwóch postaci:

- *Digital Surrogate* – to cyfrowa kopia oryginału dokumentu. Jakkolwiek TNA zachowuje oryginał dokumentu, to dla czytelników dostępna jest tylko jego kopia cyfrowa. Dla obiektu cyfrowego przechowuje się pliki macierzyste, pliki prezentacyjne, rozpoznany tekst oraz metadane METS w podstawowym zakresie;
- *Digitised Record* – to cyfrowa kopia oryginału dokumentu, spełniająca rolę dokumentu oryginalnego, który po zakończeniu digitalizacji nie będzie już przechowywany w archiwum. Przechowuje się natomiast pliki macierzyste, pliki prezentacyjne, rozpoznany tekst oraz metadane METS z bardzo wysokim poziomem szczegółowości metadanych.

Tworzenie produktów *Digitised Record* obarczone jest znacznie wyższymi wymogami technicznymi, zwłaszcza gdy chodzi o zakres i szczegółowość koniecznych do wykonania metadanych administracyjnych dotyczących proweniencji plików pochodnych, które to metadane są m.in. podstawą do udowodnienia autentyczności pełniącej rolę oryginału kopii cyfrowej.

Budząca kontrowersje rezygnacja z zachowania oryginału papierowego ma uzasadnienie. Skoro bardzo precyzyjne i szczegółowe opisanie plików obiektu cyfrowego za pomocą metadanych METS (głównie administracyjnych oraz ALTO) jest podstawą tego, by dowieść autentyczności, wiarygodności oraz kompletności dostępnego w repozytorium obiektu cyfrowego, to nic nie stoi na przeszkodzie, by obiekt taki mógł pełnić rolę oryginału. Tym samym oryginał papierowy, zwłaszcza wtedy, gdy nie spełnia warunku podlegania ochronie dziedzictwa kulturowego i którego przechowywanie pociąga za sobą określone koszty, nie musi być przechowywany w archiwum. Na przykład, w wypadku digitalizowanych czasopism powstałych w XIX w. na kwaśnym papierze można nie przechowywać oryginału papierowego. Destrukcja obiektów powstałych na kwaśnym papierze jest nieunikniona, a w jej konsekwencji dostępna będzie tylko i wyłącznie co najwyżej cyfrowa kopia obiektu. Realizowanie digitalizacji z uwzględnieniem wykonania kompletu metadanych pozwoli na stworzenie repozytoriów cyfrowych, w odniesieniu do których wykazanie wiarygodności, autentyczności nie sprawi kłopotu, a co za tym idzie będzie można stwierdzić, że ochronę dziedzictwa kulturowego przeprowadzono rzetelnie.

Metadane plików komputerowych

Metadane deskryptywne opisują zarówno oryginał obiektu, jak i jego cyfrową kopię. W przypadku publikacji zwartych metadane opisują jeden obiekt, w przypadku publikacji ciągłych – dzięki mechanizmom dziedziczenia czy też polimorfizmu – znacząca część rekordu metadanych deskryptywnych opisuje wszystkie obiekty w obrębie tej publikacji. Pozostałe kategorie metadanych

[18] The National Archives [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.nationalarchives.gov.uk/>.

opisują już wyłącznie różnorodne właściwości plików kopii cyfrowej obiektów i to bez względu na to, czy kopia taka dotyczy obiektu publikacji zwartej, czy też ciągłej. Położenie dużego nacisku na metadane inne niż deskryptywne nie ma na celu zwiększenia efektywności przeszukiwania zasobów przez biblioteki cyfrowe, bo nie do takich zastosowań metadane te zostały opracowane. Przeszukiwanie zasobów to zaledwie jedna z możliwości wykorzystania metadanych o obiekcie cyfrowym, pozostałe to przede wszystkim udostępnienie właściwości poszczególnych plików obiektu cyfrowego, określenie relacji pomiędzy plikami dotyczącymi tego samego obiektu, określenie sposobu i zakresu ich stosowania, udowodnienie autentyczności zawartości kopii cyfrowej, ułatwienie zarządzania nimi oraz weryfikowanie spójności i kompletności zgromadzonego zasobu cyfrowego. Być może właśnie dlatego, iż archiwizacja lub zarządzanie dużymi ilościami plików nie jest wprost związane z digitalizacją, brak jest szeroko stosowanych metadanych, których zasięg operowania dotyczyłby całych repozytoriów cyfrowych, a więc metadanych, które opisywałyby nie pojedynczy obiekt lub kolekcję cyfrowych obiektów publikacji ciągłej, ale wszystkie obiekty zgromadzone w danej bibliotece lub wszystkie obiekty wykonane w ramach pojedynczego projektu digitalizacji.

Prawdopodobieństwo tego, że w budowanym przez lata repozytorium cyfrowym wystąpią pliki cyfrowe o tych samych nazwach, nie jest zerowe. Rodzi to pytania czy nazwy wielokrotne, które się pojawiły, związane są z wielokrotną digitalizacją tego samego obiektu, a może digitalizacją różnych obiektów, których plikom nadano takie same nazwy, a może powieleniem posiadanych już plików i przypadkowym usunięciem innych, a może... Sytuacja taka jest dalece niekomfortową. Okazuje się, że dzięki obecnym już na rynku gotowym propozycjom – standardom metadanych EXIF oraz XMP – rozwiązanie problemu nie nastręcza większego kłopotu. Różnicę sposobu, w jaki udostępnia metadane wewnętrzne plików MS Windows7[19] i MS Windows8 w porównaniu do MS WindowsXP, ocenić można jako różnicę kilku epok informatycznych. Aby wygodnie zapoznawać się z kolekcjami plików obiektów cyfrowych zachowanych w odpowiednio powstałych gałęziach folderów, wystarczy uruchomić okno eksploratora Windows. Dotyczy to plików w wielu formatach, w tym w formatach.TIFF, .DjVu, .pdf, .JPG, .PNG, .JPG2000 a nawet.GIF, czyli niemal we wszystkich formatach plików, które stosowane są w polskich i europejskich bibliotekach. Metadane, z którymi możemy zapoznawać się np. za pomocą eksploratora Windows, to metadane po części komercyjne (np.: *tags, rating*), po części deskryptywne (niemal wszystkie atrybuty obecne w standardzie DCMES), techniczne, opisujące właściciela obiektu (np.: *Company, Project*), prawa autorskie, szeroki zestaw dat (utworzenia, pozyskania, modyfikacji, archiwizacji, ...) oraz, co bardzo istotne, metadane, które mogą być wykorzystane jako metadane operujące w obrębie całego projektu digitalizacji lub całego repozytorium cyfrowego. W tym miejscu wymienić można kilka spośród nich:

- numer pliku – unikatowy w obrębie pojedynczego projektu lub całego zasobu cyfrowego,

[19] Windows Properties [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://msdn.microsoft.com/en-us/library/windows/desktop/dd561977%28v=vs.85%29.aspx>.

- liczba plików – z których składa się publikacja wielostronicowa,
- liczba stron – publikacji wielostronicowej,
- opis pliku – np. „strona 3/8”, „strona 4/8”, co w przypadku formatów takich jak TIFF czy JPG jest bardzo użyteczne,
- źródło – czyli prawidłowa lub oryginalna lokalizacja pliku na serwerze zapisana jako ścieżka bezwzględna lub jako ścieżka względna (jeżeli zastosowano adresowanie względne),
- nazwa nośnika – na którym są (powinny być) zapisane archiwizowane pliki publikacji cyfrowej.

W przypadku propozycji LOC dotyczącej standardów dla metadanych każda odrębna tematycznie kategoria metadanych ma własną nazwę i własny standard. W przypadku producentów rozwiązań komercyjnych – w tym wypadku Microsoft – wszelkiego typu metadane, które udostępniane są za pośrednictwem Windows7 lub Windows8, to tak naprawdę jedna wielka grupa informacji o pliku cyfrowym oraz o jego zawartości. Metadane te określić można jako metadane uniwersalne (jeżeli chodzi o ich zawartość), a dokładniej, jako uniwersalne metadane wewnętrzne (zapisywane są wewnątrz pliku, którego dotyczą). Ponieważ są zapisywane w plikach, możliwość korzystania z nich nie jest determinowana odpowiednio autoryzowanym dostępem do określonej bazy danych biblioteki, lecz realizowana jest bezpośrednio, w analogiczny sposób, w jaki odczytywana jest nazwa pliku zapisanego na dysku. Dotyczy to zarówno systemu operacyjnego, jak i dowolnego oprogramowania użytkowego (np. przeglądarki metadanych *ExifTool*).

Funkcjonalność i prostotę wykonania metadanych wewnętrznych w publikacjach DjVu potwierdziła digitalizacja w ramach projektu „NUKAT – autostrada informacji cyfrowej”. Okazuje się, że w analogiczny sposób, jak w przypadku plików w formacie DjVu, integrować można uniwersalne metadane wewnętrzne w formatach plików: .TIFF, .pdf, .JPG, .JPG2000, .PNG lub .GIF (tylko dla tych formatów wykonano próby zapisu metadanych, więc nie należy zakładać, że w innych formatach plików operacje te nie będą dostępne). Do wygodnego zapoznawania się z plikami obiektów cyfrowych lub realizowaniu operacji w obrębie całego repozytorium nie ma potrzeby zapisu w każdym pliku elektronicznym kompletu metadanych. Istotnymi są jedynie podstawowe informacje opisowe oraz wyszczególnione powyżej atrybuty, którym przypisać można zasięg całego zasobu lub całego projektu digitalizacyjnego, a nie wyłącznie pojedynczego obiektu. Oczywiście nic nie stoi na przeszkodzie, by dodatkowo przewidzieć możliwość użycia atrybutów, z których dana instytucja korzystać będzie być może dopiero w przyszłości (np. atrybut „powiązany plik audio”, powiązany z cyfrową kopią czasopisma).

Uwzględniając możliwości gromadzenia metadanych, jakie oferują standardy EXIF oraz XMP, do wykonania metadanych wewnętrznych w plikach obiektów cyfrowych wytypować można niewiele ponad 30 potencjalnych atrybutów, które będą przydatne dla wszelkich operacji związanych z zarządzaniem plikami w całym zasobie cyfrowym, około 20 atrybutów wytypować można jako podstawowe. Jeżeli integralną częścią procesu digitalizacji jest wykonanie metadanych deskryptywnych, administracyjno-technicznych oraz strukturalnych (np. jako pliki metadanych METS), istotne jest, iż dodatkowe wykonanie uniwersalnych

metadanych wewnętrznych wiąże się z niewielkim nakładem pracy – uniwersalne metadane wewnętrzne zapisywane w plikach obiektów cyfrowych są jedynie efektem ponownego wykorzystania zgromadzonych już informacji o obiekcie cyfrowym.

W ramach projektu „NUKAT – autostrada informacji cyfrowej" metadane uniwersalne zapisano w około 200 000 stron zdigitalizowanych obiektów. Wykonywanie metadanych wewnętrznych dla plików macierzystych oraz prezentacyjnych nie wymaga kolejnych zmian w schematach digitalizacji. Schemat digitalizacji na rys. 7 uwzględnia operacje zapisu uniwersalnych metadanych wewnętrznych zarówno w plikach macierzystych jak i w plikach prezentacyjnych. Przykład uniwersalnych metadanych wewnętrznych, zapisanych w pliku macierzystym jednej ze stron publikacji cyfrowej, zaprezentowano na rys. 5.

Korzyści z opisywania dokumentów cyfrowych metadanymi nie tylko deskryptywnymi wydają się oczywiste. Zastanawiać natomiast może fakt stosowania uniwersalnych metadanych wewnętrznych w sytuacji, w której dostępne są również niemal te same metadane przechowywane pod postacią tekstowych plików metadanych METS lub wprost – w bazach danych. Aby metadane różnych kategorii tematycznych gromadzić pod postacią metadanych METS, konieczne jest rozszerzenie posiadanej bazy sprzętowej oraz nabycie specjalistycznego oprogramowania. Pomijając znaczące i konieczne koszty związane z wdrożeniem takiej decyzji, efektywna obsługa takiego oprogramowania wymaga określonej liczby szkoleń, wykonanych prób i testów. W przypadku uniwersalnych metadanych wewnętrznych możliwość ich zapisu w cyfrowych plikach publikacji realizowana może być za pomocą nieskomplikowanych w obsłudze aplikacji. Sposób ich przechowywania nie wpływa na stan i postać eksploatowanych baz danych, rozpoczęcie ich gromadzenia dla już stosowanych systemów informatycznych, jest niezauważalne. Ponadto zyskuje się możliwość wyeksportowania ich w przyszłości do takiej postaci, która umożliwi ich zapis w postaci plików metadanych METS.

Mając na uwadze sposób, w jaki zrealizowana została digitalizacja w ramach projektu „NUKAT – autostrada informacji cyfrowej" oraz wcześniej wypracowane standardy pracy zespołu pracowników BUW, stwierdzić można, że analogicznie jak w przypadku rozpoczęcia tworzenia biblioteki cyfrowej e-bUW, kiedy to oczywistą koniecznością było gromadzenie metadanych deskryptywnych dla powstających kopii cyfrowych obiektów, tak i realizacja wyżej wymienionego projektu nie tylko potwierdziła, ale i udowodniła, że dla potrzeb zarządzania stale narastającymi zasobami cyfrowymi konieczne jest wykonywanie i przechowywanie również metadanych administracyjno-technicznych oraz metadanych strukturalnych. Ponadto okazało się, że zachowywanie zgromadzonych metadanych pod postacią metadanych wewnętrznych skutkuje tym, iż korzystanie z nich możliwe jest nie tylko za pośrednictwem prostych w użyciu i popularnych narzędzi informatycznych, ale również za pomocą niektórych opcji systemu operacyjnego (np. MS Windows7). Tym samym, z zawartością takich metadanych zapoznawać mogą się nie tylko specjaliści z zakresu informatyki, ale także i bibliotekarze czy też archiwiści, dla których komputer jest jedynie narzędziem, które powinno ułatwiać wykonywanie codziennej pracy.

Bibliografia

1. Adobe's Extensible Metadata Platform (XMP), Part 1, Data model, Serialization, and Core Properties. W: Adobe XMP Developer Center. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/cs6/XMPSpecificationPart1.pdf>.
2. Adobe's Extensible Metadata Platform (XMP), Part 2, Additional Properties. W: Adobe XMP Developer Center. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/cs6/XMPSpecificationPart2.pdf>.
3. Adobe's Extensible Metadata Platform (XMP), Part 3, Storage in Files. W: Adobe XMP Developer Center. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/cs6/XMPSpecificationPart3.pdf>.
4. Analyzed Layout and Text Object (ALTO). [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.loc.gov/standards/alto/>.
5. Burchard M., Wołodko A.: *NUKAT – autostrada informacji cyfrowej*. W: *Infobazy 2011 – Nauka, Projekty Europejskie, Społeczeństwo Informacyjne: materiały VI krajowej konferencji naukowej. Gdańsk – Sopot, 5-7 września 2011*. Gdańsk 2011, s. 259-265.
6. Bednarek G.: *Integracja i wykorzystanie metadanych w publikacjach DjVu*. W: *Polskie Biblioteki Cyfrowe 2008. Materiały z konferencji zorganizowanej w dniach 24-25 listopada 2008 r. przez: Bibliotekę Kórnicką PAN, Poznańską Fundację Bibliotek Naukowych, Poznańskie Centrum Superkomputerowo-Sieciowe*. Pod red. C. Mazurka, M. Stroińskiego, J. Węglarza. Poznań 2009, s. 89-105. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://lib.psnc.pl/Content/209/pbc-11-Bednarek.pdf>.
7. CIPA DC-008-Translation-2010. Exchangeable image file format for digital still cameras: Exif Version 2.3. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.mets-alto.eu/Metadata/Standards/Specs/EXIF/DjVu_File.html>.
8. Metadata Encoding and Transmission Standard (METS). W: The Library of Congress. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.loc.gov/standards/mets/>.
9. Metadata Object Description Schema (MODS). W: The Library of Congress. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.loc.gov/standards/mods/>.
10. NISO Technical Metadata for Digital Still Images (MIX). W: The Library of Congress. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.loc.gov/standards/mix/>.
11. Potęga J., Wróbel A.: *The Dublin Core Metadata Element Set, Ver. 1.1 a potrzeby i oczekiwania bibliotekarzy cyfrowych – analiza przypadków*. W: *Polskie Biblioteki Cyfrowe 2010. Materiały z konferencji zorganizowanej dnia 9 grudnia 2009 r. przez: Bibliotekę Kórnicką PAN, Poznańską Fundację Bibliotek Naukowych, Poznańskie Centrum Superkomputerowo-Sieciowe*. Pod red. C. Mazurka, M. Stroińskiego, J. Węglarza. Poznań 2010, s. 71-78. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://lib.psnc.pl/Content/367/08-Potęga-ER.pdf>.
12. The PREMIS Data Dictionary for Preservation Metadata. W: The Library of Congress. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.loc.gov/standards/premis/>.
13. *Standardy w procesie digitalizacji obiektów dziedzictwa kulturowego*. Pod red. G. Płoszajskiego. Warszawa 2008. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://bcpw.bg.pw.edu.pl/publication/1113>.
14. Use of METS and ALTO in the Australian Newspapers Digitisation Program. W: The National Library of Australia. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.nla.gov.au/ndp/project_details/documents/ndpuseofmetsandaltoJune2010doc.pdf>.
15. Warunki wpływu obiektów cyfrowych do BN 2011-10-07. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.bn.org.pl/download/document/1332343355.pdf>.
16. Warunki wpływu obiektów cyfrowych do BN metadane 2012-03-21. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.bn.org.pl/download/document/1332406902.pdf>.

17. Wróbel A.: *Opracowanie dokumentów w bibliotece cyfrowej e-bUW*. W: *Z Problemów Bibliotek Naukowych Wrocławia, z. 10: III Wrocławskie Spotkania Bibliotekarzy*. Pod red. H. Szarskiego, D. Dudziak. Wrocław 2011, s. 321-329. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.dbc.wroc.pl/dlibra/docmetadata?id=12277>.
18. Yale University Library Best Practices. W: Digital Production & Integration Program (DPIP) Yale University Library's. [online]. [dostęp: 2.05.2013]. Dostępny w World Wide Web: <http://www.library.yale.edu/dpip/bestpractices/>.

Summary

Discussion on metadata in Polish digital libraries focused mostly on descriptive metadata. The article presents benefits of application of different metadata standards in a digital library. This enables collection of numerous, different information of a digital objects (administrative and technical metadata), which have not been collected in Polish libraries so far. Moreover, the authors prove, that different methods of metadata collection (saving in a digital object, saving in an another file) complement each other, and offer an additional protection against data loss. Some of the solutions presented in the article have been implemented during the project "NUKAT – Autostrada Informacji Cyfrowej" by the digital library e-bUW.